The Ohio State University Press

Geographical Analysis

An International Journal of Theoretical Geography

January 2003 ■ Vol. 35 ■ No. 1

Geographical Analysis

An International Journal of Theoretical Geography

Editor

Alan T. Murray
The Ohio State University

Associate Editors

A. Stewart Fotheringham, *University of Newcastle*
Daniel A. Griffith, *Syracuse University*
Mei-Po Kwan, *The Ohio State University*

Consulting Editors

Luc Anselin, *University of Illinois*
Marc P. Armstrong, *University of Iowa*
Richard L. Church, *University of California, Santa Barbara*
C. Cindy Fan, *University of California, Los Angeles*
Manfred M. Fischer, *Vienna University of Economics and Business Administration*
Mark Gahegan, *Penn State University*
Arthur Getis, *San Diego State University*
Michael F. Goodchild, *University of California, Santa Barbara*
Suchi Gopal, *Boston University*
Kingsley E. Haynes, *George Mason University*
Geoffrey J. D. Hewings, *University of Illinois*
Ron Johnson, *Bristol University*
Pavlos Kanaroglou, *McMaster University*
Nina S-N. Lam, *Louisiana State University*
James P. LeSage, *University of Toledo*
Sara McLafferty, *University of Illinois*
Harvey Miller, *University of Utah*
Peter Nijkamp, *Free University*
Morton O'Kelly, *The Ohio State University*
Atsuyuki Okabe, *University of Tokyo*
Kavita Pandit, *University of Georgia*
Jonathan D. Phillips, *University of Kentucky*
David Plane, *University of Arizona*
Serge Rey, *San Diego State University*
Bruce Rhoads, *University of Illinois*
Peter Rogerson, *University at Buffalo, The State University of New York*
Yukio Sadahiro, *University of Tokyo*
Jean-Claude Thill, *University at Buffalo, The State University of New York*
Elizabeth Wentz, *Arizona State University*

Executive Committee

Lawrence A. Brown (chair), *The Ohio State University*
Reginald G. Golledge, *University of California, Santa Barbara*
Randall W. Jackson, *West Virginia University*
Edward J. Malecki, *The Ohio State University*
Joel L. Morrison, *The Ohio State University*

Geographical Analysis

An International Journal of Theoretical Geography

Volume 35 **Number 1** **January 2003**

Articles

Evaluating an Entropy-Maximizing Solution to the Ecological Inference Problem: Split-Ticket Voting in New Zealand 1999
Ron Johnston and Charles Pattie 1

Measuring Impacts of Condition Variables in Rule-Based Models of Space-Time Choice Behavior: Method and Empirical Illustration
Theo Arentze and Harry Timmermans 24

Urban Land Use for Transport Systems and City Shapes
Francesca Medda, Peter Nijkamp, and Piet Rietveld 46

Exploring the Relationship Between River Bank Erosion and Geomorphological Controls Using Geographically Weighted Logistic Regression
Peter M. Atkinson, S. E. German, D. A. Sear, and M. J. Clark 58

Measuring Accessibility for People with a Disability
Richard L. Church and James Marston 83

Copyright 2003 The Ohio State University

ISSN 0016-7363. Published quarterly by The Ohio State University Press, 1070 Carmack Road, Columbus, Ohio 43210-1002, USA. Eavon Lee Mobley, Managing Editor; Jodie Moffett, Subscriptions Manager. Periodicals postage paid at Columbus, Ohio, and at additional mailing offices. Copyright © 2003 by The Ohio State University. *Postmaster*: Send address changes to: *Geographical Analysis*, The Ohio State University Press, 1070 Carmack Road, 180 Pressey Hall, Columbus, Ohio 43210-1002, USA.

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials. ANSI Z39.48-1992.

2003 Subscriptions. *Geographical Analysis* is published in January, April, July, and October. Annual subscription rates: libraries and institutions, $135.00; individuals, $47.00; students, $25.00. Orders outside the U.S.: add $12.00 postage. Canadian orders add 7% GST (GST 123995904). Orders should be addressed to The Ohio State University Press, 1070 Carmack Road, Columbus, Ohio 43210-1002, USA. Claims for missing issues should be made within four months of issue date. Allow six weeks for address change. Please contact the publisher for availability of back issues. Reprints of individual articles, reprints of out-of-print issues, and microfilms of back volumes are available from University Microfilms International, 300 North Zeeb Road, Ann Arbor, Michigan 48106, USA. On the Web at www.umi.com.

Editor's Introduction

I am truly honored to be selected as the eighth editor of *Geographical Analysis*. At the same time, these are very big shoes to fill, as pointed out to me by colleagues. The journal is now beginning volume 35, and continues to publish 4 issues each year. *Geographical Analysis (GA)* has come to be the premier journal focusing on geographical theory, modeling, and quantitative methods. This is certainly a noteworthy achievement, and sustaining this stature within the discipline will depend on the submission and publication of rigorous and innovative research.

The journal was started back in 1969 in the Department of Geography at The Ohio State University. It has been edited by faculty in this department, most recently, since 1991. Rough estimates show that there have been over 320 submissions since 1990, out of which approximately 250 have been published. Reviews for these submissions have been provided by some 400 different scholars. The rate of acceptance over this period is about 50%. As the number of publications per year has remained somewhat consistent, it would appear that there have been well over 700 published articles in *GA* requiring the attention of perhaps more than 1000 experts spanning a range of specialties and disciplines. Viewed from this perspective, the impact of *GA* has been substantial. This undoubtedly reflects the relatively high impact factor rating of the journal (1.024 according to the 2000 rating of the Journal Citation Reports produced by the Institute of Scientific Information).

The journal, and all that is associated with it, is the result of quality submissions that have been subject to the highest standards of peer review. In addition, the past editor Morton O'Kelly (1996–2002) and those that have served before him (Les King 1969–73, Reg Golledge 1974–78, Ross MacKinnon 1979–84, Bill Clark 1985–87, Mike Goodchild 1988–90, and Emilio Casetti 1991–95) are to be commended for effectively guiding and positioning the journal, resulting in the high regard and respect for *GA* in Geography and associated disciplines. On behalf of the previous editors, I would like to also thank past Associate Editors (10 thus far, not including the three new additions) and Consulting Editors (94 thus far, not including the 15 new additions) for their invaluable service contributions to the journal:

Associate Editors

R.G. Golledge (1/71–1/74)
D.R. Reynolds (1/74–10/78)
J.N. Rayner (7/77–10/78)
Emilio Casetti (1/79–10/84)
Morton O'Kelly (1/85–10/95)
R.J. Bennett (1/85–10/87)
John P. Jones III (1/91–10/96)
Robert Haining (1/96–10/02)
Kavita Pandit (4/97–10/02)
Alan T. Murray (7/01–10/02)

Consulting Editors

H.J. McCarty (1/69–10/74)
J.D. Nystuen (1/69–10/74)
G. Tornqvist (1/69–10/74)
W. Warntz (1/69–10/74)
A. Wrobel (1/69–10/74)
B.J.L. Berry (1/69–10/78)
G.H. Dury (1/69–10/78)
P.R. Gould (1/69–10/78)
A.L. Mabogunje (1/69–10/78)
D.F. Marble (1/69–10/78)
J. Wolpert (1/69–10/78)
R.H.T. Smith (1/69–10/79)
P. Haggett (1/69–7/81)
L. Curry (1/69–10/87)
D.W. Harvey (7/69–10/78)
G. Olsson (7/69–10/87)
Y.V. Medvedkov (1/69–10/78)
Z. Chojnicki (1/74–10/78)
O. Warneryd (1/74–10/78)
F. Lukermann (1/74–10/79)
J.C. Hudson (1/74–10/82)
L.J. King (1/74–10/90)
W.R. Tobler (1/74–10/95)
B.J. Garner (7/76–10/84)
Jeffrey P. Osleeb (10/77–10/78)
D.R. Reynolds (1/79–7/81)
S. Gale (1/79–10/82)
R.S. Jarvis (1/79–10/84)
B. Marchand (1/79–10/84)
N. Alao (1/79–10/87)
R.G. Golledge (1/79–10/87)
Y.Y. Papageorgiou (1/79–10/87)
P. Korcelli (1/79–10/90)
John Rayner (1/79–10/00)
A.J. Arnfield (1/80–10/84)
E.S. Sheppard (1/80–10/90)
G. Leonardi (10/81–10/84)
J. Thornes (10/81–10/84)
Robert Haining (10/81–10/95)
John Odland (10/81–10/02)

D.A. Griffith (1/83–10/90)
J.S. Pipkin (1/83–10/90)
J.M. Batty (1/85–10/87)
H. Beguin (1/85–10/87)
R.D. MacKinnon (7/78–10/87)
D.F. Batten (1/85–10/90)
E.G. Moore (1/74–10/90)
H. Timmermans (1/85–10/90)
Manfred M. Fischer (1/85–present)
R.J. Bennett (1/79–10/90)
A. Ghosh (1/88–10/90)
Y. Leung (1/88–10/90)
D.M. Mark (1/88–10/90)
D. Pumain (1/88–10/90)
C.S. ReVelle (1/88–10/90)
T.E. Smith (1/88–10/90)
J.-F. Thisse (1/88–10/90)
David Plane (1/88–present)
M.J. Webber (1/74–10/90)
William A.V. Clark (1/79–10/02)
D.S. Dendrinos (1/91–10/95)
J.C. Knox (1/91–10/95)
N. Wrigley (1/88–10/95)
Anthol D. Abrahams (1/91–10/02)
Lawrence A. Brown (1/91–10/02)
A. Stewart Fotheringham (1/91–10/02)
Howard L. Gauthier (1/91–10/02)
Pat Gober (1/91–10/02)
Luc Anselin (1/91–present)
Michael F. Goodchild (1/91–present)
Kingsley E. Haynes (1/91–present)
Geoffrey J.D. Hewings (1/91–present)
Arthur Getis (7/91–present)
C. Werner (7/76–10/95)
Nina S-N. Lam (10/93–present)
Kelvyn Jones (4/94–10/02)
Ayse Can Talen (1/96–10/02)
Emilio Casetti (1/96–10/02)
Helen Couclelis (1/96–10/02)
Adrian Esparza (1/96–10/02)
Hee-Yeon Lee (1/96–10/02)
Brigitte Waldorf (1/96–10/02)
Cort Willmott (1/96–10/02)
Randall W. Jackson (1/96–10/02)
Harvey Miller (1/96–present)
Atsuyuki Okabe (1/96–present)
Kavita Pandit (1/96–present)
Sergio Rey (1/96–present)
Jean-Claude Thill (1/96–present)
John P. Jones III (10/97–10/02)
Deborah Bryan (1/98–10/02)
Mei-Po Kwan (10/00–10/02)
James P. Lesage (10/00–present)
Brian Mikelbank (1/01–10/02)

(I should note that some of the above service periods for Consulting Editors have not been consecutive)

We have assembled an exceptional board to help maintain the high standards of the journal as well as to represent the changing nature of geographical theory. I am grateful to the Associate Editors, Consulting Editors and Executive Committee (all acknowledged on the inside cover of each journal issue) for their dedication to and support of the journal. I plan to rely extensively on the knowledge and expertise of the editorial board in guiding the journal over the coming years.

As editor, a major objective will be to continue to maintain, and hopefully build upon, the success of past editors of *GA*. We are soliciting and encouraging submissions that contribute to the continued development of theoretical geography. We hope that published research in the coming years continues to build upon, extend and evolve theory and methods applied in a wide variety of geographical contexts, both physical and human.

The journal will entertain "Research Notes and Comments" submissions, which appear to have subsided in 1993, and "Methodological Discussions", explored in 1983. Further, we hope to publish one innovative "Special Issue" each year as has been done sporadically in the past ("Spatial Temporal Dynamics" in 1981; "Location Analysis and Modeling" in 1990; "Spatial Modeling" in 1992; and "The Multinodal Metropolis" in 1993). By the way, copies of these and all other back issues of *GA* have now been converted to electronic format (http://muse.jhu.edu/journals/geographical_analysis/).

An operational goal is to streamline the reviewing process, to the extent this is possible, by encouraging electronic submission of papers (currently we are accepting either MS Word or PDF formatted submissions). This will facilitate the distribution process, both to the journal and to referees. A source of pride is certainly that *GA* provides timely and thoughtful reviews, so we hope to maintain this record.

The journal is only as good as the submissions it receives and the subsequent quality refereeing that is applied to these submissions. I look forward to facilitating the publication of the very best of geographical theory, modeling, and quantitative methods in *GA* in the coming years.

Alan Murray
Editor, *Geographical Analysis*
ga@geography.ohio-state.edu

Ron Johnston
Charles Pattie

Evaluating an Entropy-Maximizing Solution to the Ecological Inference Problem: Split-Ticket Voting in New Zealand, 1999

For geographers, one important aspect of the ecological inference problem relates to the analysis of spatial variations in individual behavior. Obtaining estimates of this behavior, in the absence of direct data, is often difficult, and the "solution" to the ecological inference problem propounded by G. King is not relevant in all circumstances. An alternative, using a different approach, has been used for some time in electoral studies but has lacked "real" data against which to assess the accuracy of its estimates. The availability of such data for New Zealand's 1999 general election allows such an assessment to be made—with very favourable results. The estimates are then used to test hypotheses regarding the spatial variation in split-ticket voting, again with considerable success.

The problems of ecological inference and fallacies present major constraints to the analysis of spatial variations in a wide range of individual behaviors. In many situations, no data are available with which to map such (supposed?) variations, whereas in others the only available data refer to population aggregates, analysis of which provides circumstantial evidence only of the spatial variations of interest.

Over the decades since the nature of these problems was first recognized, social scientists have advanced a variety of methods for tackling them. In 1997, however, King (1997) claimed to have solved the problem with a procedure that produces estimates of the unknown quantities and thereby provides the needed information for analyzing spatial variations. That procedure has drawbacks, however, particularly with regard to multi-state situations (i.e., where the behavior being considered is more than binary in the number of states), hence continued work on alternatives (as in King, Rosen, and Tanner 1999; Rosen et al. 2001). Furthermore, although King's

The split-ticket voting matrix in Table 1 was produced by the New Zealand Electoral Study 1999 (http://www.nzes.org/1999/splitvote.html). We are grateful to Professor Jack Vowles of the Department of Politics, University of Auckland, for this and other data, and the New Zealand Electoral Commission for providing the spending data.

Ron Johnston is a professor in the School of Geographical Sciences, University of Bristol. Email: r.johnston@bristol.ac.uk. *Charles Pattie is a professor in the Department of Geography, University of Sheffield. E-Mail:* c.pattie@sheffield.ac.uk

Geographical Analysis, Vol. 35, No. 1 (January 2003) The Ohio State University
Submitted: March 21, 2002. Revised version accepted: June 27, 2002.

method has been employed to produce estimates of spatial variations in patterns of behavior—as in Burden and Kimball's (1998) on ticket-splitting at U.S. federal elections—the focus has been almost entirely on estimating the internal cells of 2 x 2 matrices, with little or no reference to those cells in the set of sub-matrices from which the overall estimates are derived. (See, for example, Kousser 2001, and other papers in his special issue on ecological inference in historical studies.)

One of those alternatives is entropy-maximizing (*EMax*), a procedure first introduced to geographers more than three decades ago (Wilson 1967, 1970). It has been adapted to tackle particular forms of the ecological inference problem—mainly in the study of voting behavior (as first set out in Johnston and Hay 1982, 1983; see also Johnston, Pattie, and Allsopp 1988). One drawback with its application to date, however, has been the lack of data against which to check the accuracy of its estimated values. This paper reports on an application—split-ticket voting at New Zealand's 1999 general election—for which excellent data to provide such a check are available. It first establishes the goodness-of-fit between the actual and estimated spatial variation in the variables of interest, and then uses the estimated values to analyze the nature and co-variates of that variation.

THE PROBLEM

The nature of the geographical-ecological inference problem is illustrated by Figures 1 and 2.

The Empty-Cell Situation

Figure 1 shows the situation in which the analyst knows nothing of the link between the two variables involved—what we term the *empty-cell situation*. For each of a number of areas, we know the sum of individuals for each row and column of the data matrices (indicated by the letter K in the figure), but have no information about the values for their internal cells (indicated by the letter U). Therefore any analyses of spatial variations can address the question of "are there differences in the size of the cell values from place to place?" through studies of the relevant row and column sums only—the so-called method of ecological correlation. Similarly, we know only the row and column totals for the summed matrix, shown on the right-hand side of the figure (i.e., the sums of all the individual area matrices).

As an example of this situation, assume that the rows in each matrix in the column refer to individuals' social class and the columns to the party they voted for at a given election. Regressing, say, the percentage of the votes cast for party I against the percentage of the population in class A may provide circumstantial evidence of the strength of support for party I among class A: if there is a strong positive relationship, then the more members of class A in an area the more votes there are for party I—from which we may draw the ecological inference that members of class A tend to vote for party I.

But that is an inference only; it could be quite feasible to find relatively high correlations between the two variables at the ecological scale from regressions involving the row and sum columns, but only small percentages of the row totals in the relevant diagonal cell of the data matrix. Thus party I may have obtained 40% of the votes cast in an area with 45% of its population in class A, and 60% of the votes in another area where 55% are in class A. But the proportion of those in class A who voted for party I may vary considerably across the areas. Because party I tends to perform well in areas where class A is large does not necessarily imply that party I gets much of its support from class A, or that most members of class A tend to vote for party I.

This issue of reconciling aggregate and individual patterns—is one aspect of the widely appreciated ecological fallacy: can we infer patterns at one scale (that of the

individual) from those at another (the aggregate)? Over the five decades since Robinson (1950) and Goodman (1959) first raised the issue, social scientists have focused a great deal of attention to answering this question.[1] One of the most recent attempts—which has attracted a great deal of attention (not least among geographers: see Sui et al. 2000)—was that of Gary King in his book *A Solution to the Ecological Inference Problem* (King 1997).

King's particular purpose was to estimate the cell values for the summed matrix, shown on the right-hand side of Figure 1 (in large part because he was addressing the problems of research trying to produce estimates of those cell values for court cases that were reliant on estimates of various ecological relationships, such as the proportion of blacks and whites who vote at U.S. elections). Using a combination of two previously suggested procedures (the method of bounds and the method of moments: see Kousser 2001; Lewis 2001), his method produced what he claimed were more reliable estimates than those produced by others—and also more "sensible" in that because the method of bounds was involved it was impossible to produce nonsense estimates (such as that 105% of those who voted for party I at the first election did so again at the second).

King's focus was on the cells of the summed matrix but one of the by-products of his method is that it also produces estimates of the cell values in each of the matrices for the separate areas so that, for example, one could obtain estimates not only of the percentage of those voting for I among class A for the entire area (the right-hand matrix), but also for each of its constituent sub-areas (the left-hand matrices). This has allowed analyses of spatial variations in such estimates across the sub-areas, as in Burden and Kimball (1998) work on split-ticket voting at the 1988 U.S. federal elections and O'Loughlin's (2000) on voting for the Nazi Party in 1930s Germany.

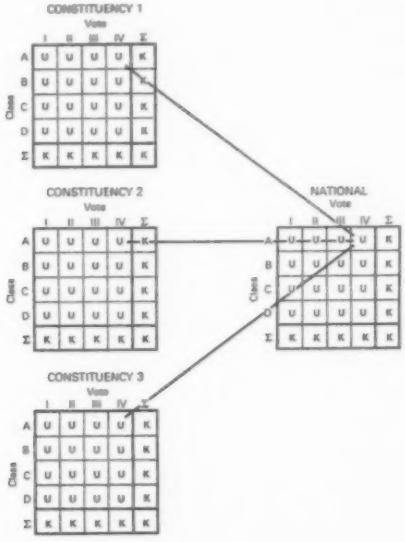

FIG. 1. The Ecological Inference Problem: the Empty-cell Situation.

1. For a recent study of the issue, see Gelman et al. (2001)

King's book has attracted a great deal of attention because of its claims to solve a major problem in quantitative social science—though its qualities have not been universally accepted by critics (see, for example, Cho 1998; Freedman, Klein, and Ostland 1998; and the reply of King (1999). For geographers interested in spatial variations in the cell values of ecological matrices, it offers substantial possibilities. But it also has drawbacks, two of which are relevant to the current discussion: 1) it is basically only applicable to 2 x 2 matrices, and applications to larger problems (such as the 4 x 4 matrices shown in Figure 1) are fraught with difficulty; and 2) it doesn't allow for constraints other than those of the row and column totals in each sub-area. For situations involving both of these, an alternative approach to estimating spatial variations is desirable.

The Sum-of-Cells Situation

Such an alternative approach has been available for some time (Johnston and Hay 1982, 1983), though the accuracy of its estimates has not been extensively evaluated. King (1997) was able to evaluate the accuracy of his estimates—at the global level of the matrix on the right-hand side of Figure 1—because he had the relevant data. This section sets out that alternative method—what we term the *sum-of-cells situation:* a later section provides by far the strongest evidence yet available of its precision and accuracy.

The nature of this situation is exemplified in Figure 2, which, like Figure 1, gives the known (K) and unknown (U) values in the same ecological problem. The clear difference between them is that in Figure 2 the internal cell values are known for the global matrix (on the right-hand side). These provide extra constraints to the estimating procedure, additional to the row and column sums employed by King. Whereas King's ecological inference (*EI*) method answers the question, What is the most likely value in each cell in each matrix (sub-area and global), given the constraint that those cell values must sum to the relevant row and column totals?

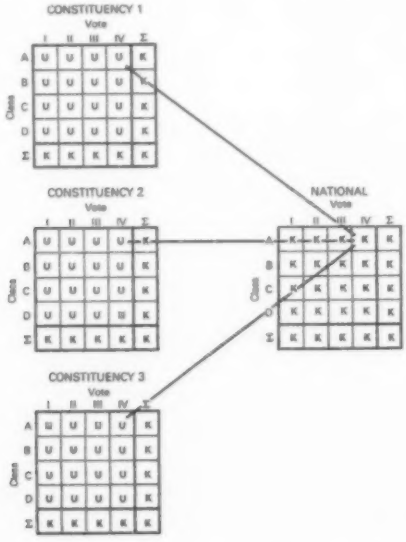
CONSTITUENCY 1
Vote
I II III IV Σ
A U U U U K
B U U U U K
C U U U U K
D U U U U K
Σ K K K K K
Data
CONSTITUENCY 2
Vote
I II III IV Σ
A U U U U K
B U U U U K
C U U U U K
D U U U U K
Σ K K K K K
Data
CONSTITUENCY 3
Vote
I II III IV Σ
A U U U U K
B U U U U K
C U U U U K
D U U U U K
Σ K K K K K
Data
NATIONAL
Vote
I II III IV Σ
A K K K K K
B K K K K K
C K K K K K
D K K K K K
Σ K K K K K
Data

FIG. 2. The Ecological Inference Problem: the Sum-of-cells Situation.

The question addressed with the procedure employed here (*EMax*) is What is the most likely value in each cell of each sub-area matrix, given the constraints that those cell values must sum not only to the relevant row and column totals but also to the relevant cell total in the global matrix (the face sum in a three-dimensional data cube)? Thus, to return to the example, whereas King's *EI* method estimates the global number of party I's supporters who are also in class A, and as a by-product obtains estimates of spatial variations in that figure across all of the sub-area matrices, *EMax* estimates the latter given knowledge of the global sum. It fills the empty cells in such a way that they sum to the known global values: *it estimates geographical variations in a known rather than an unknown quantity*.

The *EMax* procedure was introduced to geographers, for example, by Alan Wilson (1970) in his work on transport flows. Its adaptation to problems in electoral geography has been set out on a number of occasions (as in Johnston and Pattie 2000).[2] *EMax* produces the maximum-likelihood estimates for each of the unknown cell values, given the constraints. Using the example in Figure 2:

- the cell values in each row of each sub-area matrix must sum to the known row total for that matrix;
- the cell values in each column of each sub-area matrix must sum to the known column total for that matrix; and
- the values for each cell across all sub-area matrices must sum to the value of that cell in the global matrix.

Estimates of the number of party I's voters in class A in each sub-area are therefore consistent not only with the pattern of voting and class composition there but also with the total number of party I voters in class A across all sub-areas.

Applying the EMax *Procedure*

The nature of the *EMax* procedure has been set out in a number of papers (e.g., Johnston and Pattie 2000), and so will not be reproduced here. It employs an iterative procedure to identify the distribution of individuals across the cells in the sub-area matrices that is most likely to occur (i.e., the maximum-entropy solution) given the constraints of the row, column, and face sums of the three-dimensional data cube. Those are the maximum likelihood estimates, the most likely distribution given the absence of any other constraints. Fitting involves using the iterative proportional fitting procedure first set out by Deming and Stephan (1940; see also Mosteller 1968).

THE ACCURACY OF *EMAX* ESTIMATES

Although *EMax* has been used for some two decades to produce estimates of the geography of a variety of aspects of voting behavior, one of the drawbacks of these applications has been the lack of corroborative evidence regarding the accuracy of the estimates. (Until recently, only one data set was identified which allowed even limited testing: Johnston, Hay, and Rumley 1983, 1984.) Analysts (and their readers) have had to take on trust that the estimates produced are reliable—i.e., are accurate representations of the "real" (unknown) values—and the veracity of analyses of the geographies so revealed have similarly been based on that trust. (One comparative study of the merits of various attempts—pre-King—to resolve the *EI* problem did report favorably on it —Cleave, Brown, and Payne 1995; and it was applied by one other researcher: Berg 1988.) The availability of a new data set allows a much more detailed

2. Although this literature seems to be largely unknown among geographers: Johnston and Pattie (2001).

check on the claims than has heretofore been possible; it provides the best-ever test of the accuracy of the entropy-maximizing estimates.

MMP Elections in New Zealand

After two national referendums, in 1996 New Zealand held its first general election to the unicameral national Parliament (the House of Representatives), using a proportional representation electoral system—Mixed-Member Proportional (MMP). (Previous elections had used the first-past-the-post system in single-member constituencies.) This system involves what are, in effect, two elections held contemporaneously. Electors have two votes: 1) a *constituency vote*, to elect a single MP to represent their local constituency, in a first-past-the-post contest—at the 1999 general election there were 67 such constituencies; and 2) a *list vote*, to elect MPs from a particular party—at the 1999 general election, there were 53 such MPs.
The procedure then works as follows:

1. 67 MPs are elected to the House of Representatives from the single-member constituencies;
2. The percentage of the list votes won by each party is calculated, to indicate the overall percentage of seats in the House to which each is entitled: entitlement is constrained to parties which either get 5% of the list votes nationally or win one or more of the 67 constituency seats;
3. The party composition of the House overall is thus determined (i.e., the 120 seats), based on the percentages calculated for the eligible parties under step 2;
4. For each party, the number of seats won in the single-member contests (step 1) is subtracted from its total allocation of seats (step 3), and the remaining seats to which it is entitled are allocated from the list.

The list votes are crucial to determining the composition of Parliament, and so to the nature of the government. Members of the electorate should thus cast their list vote for their preferred party (what is known as their sincere vote): the more votes it gets in this contest, the more seats it will get in Parliament, irrespective of its number of victories in the constituency contests. Their constituency vote need not be for a candidate from that party, however: they may decide to vote for a candidate from another party—perhaps because of her/his individual characteristics or because they believe a constituency vote for the candidate of their preferred party in the list contest would be wasted because he/she has no chance of winning there. There may thus be some split-ticket voting across the two contests, with members of the electorate voting for different parties in each.

How much split-ticket voting is there—and does it vary geographically, across the 67 constituencies? The answer to the first of these questions is given by the Electoral Commission, which assembles data from the voting returns on the volume of split-ticket voting: in 1996, 37% of the electorate voted a split ticket, as did 35% in 1999. The percentage that voted a split ticket varied across the constituencies, however: from 31 to 77 in 1996 and from 28 to 73 in 1999. (On the 1996 election and the estimates of split-ticket voting derived then, see Johnston and Pattie 1999.)

If people vote a split ticket, what is the pattern of the splits—how many vote for each pair of parties at the two contests? And do these percentages vary across the constituencies? The answer to the first of these questions is provided by sample surveys of the electorate. The 1999 New Zealand Election Study conducted a large survey of 5382 voters, giving the split-ticket voting matrix in Table 1. This shows, for example, that of the 428 who voted for the Alliance in the list contest (144) 34% also voted for its candidate in the constituency contest, with a further 209 (49%) switching their support to the Labour Party candidate, and only small numbers splitting their ticket to support another party.

TABLE 1
The New Zealand Election Study Split-ticket Voting Matrix, 1999

List vote	Constituency vote							
	A	ACT	CC	G	L	N	NZF	O
A	144	3	3	17	209	16	14	22
ACT	4	93	3	5	32	226	1	3
CC	2	2	56	1	19	30	2	12
G	21	10	1	64	123	21	9	35
L	200	19	8	78	1750	89	43	134
N	18	110	10	13	145	1189	21	85
NZF	5	1	2	6	49	22	136	24
O	14	8	10	11	98	55	9	40

NOTE: Key to parties: A—Alliance; ACT—ACT; CC—Christian Coalition; G—Green; L—Labour; N—National; NZF—New Zealand First; O—Other.

The second question can be addressed using *EMax*. For each of the 67 constituencies used at the 1999 general election, we know—from Electoral Commission returns—the number of votes for each candidate in the constituency contests and for each party in the list contest; these form the column and row totals in the sub-area matrices in Figure 2. The national split-ticket matrix (Table 1), grossed up from the sample size to the total electorate, provides the third set of constraints—the global split-ticket matrix. With these three, we used *EMax* to estimate the unknown cell values for each of the 67 constituency matrices.

The Accuracy of EMax Estimates: Party Loyalists by Constituency

How accurate are those estimates? The Electoral Commission publishes the total number of straight-ticket votes (i.e., votes for the same party at both contests) in each constituency—these are the sum of the diagonal values (top-left to bottom-right) in each constituency matrix. In addition, it produces counts for the number of straight-ticket votes for each party—i.e., for the individual diagonal values in each constituency. This gives us the actual values for eight of the 64 cell values in each of the 67 matrices, against which the accuracy of the *EMax* estimates can be checked.

Scatter-plots of the relationships between the actual and estimated number of loyalists for each party and for the total are given in Figures 3 through 11. In all but one of these, the closeness of the two values in every constituency—and thus of each constituency to the best-fit linear regression line—is very substantial indeed. Even with those parties that performed much better in a small number of constituencies than they did in the great majority (Alliance, New Zealand First, Green) the fit is very good across all constituencies (and there is no undue leverage on the regression caused by the "outliers"). The estimates and the actual values are extremely similar in all cases.

The regression parameters in Table 2 confirm the visual conclusions drawn from the nine figures. Excluding the one deviant case (other parties: see below), the average value of r^2 for the seven individual parties is 0.961 (which is also the value of r^2 for all voters together—the total graph in Figure 11), and the standard deviation is only 0.03, showing a very tight clustering about the mean—all providing clear evidence of very close fits. Further, the a values are all close to zero and the b coefficients close to 1.0, indicative of a virtual 1:1 ratio between the actual and estimated values.

The only exception to this overall pattern of very close fits between estimated and actual straight-ticket voting is for "other" parties. The reasons for this are straightforward. A considerable number of small parties stood in the list contests, but many of them fielded at most just a few candidates for the constituency contests (several

TABLE 2
Regressions of Estimated Against Actual Numbers of Party Loyalists

	a	b	r^2
Labour	629	0.913	0.984
National	492	0.923	0.980
Alliance	−38	0.886	0.918
New Zealand First	39	0.906	0.988
ACT	13	0.873	0.959
Green	−51	0.937	0.963
Christian Coalition	−27	1.109	0.934
Other	137	0.819	0.402
Total	1781	0.888	0.961

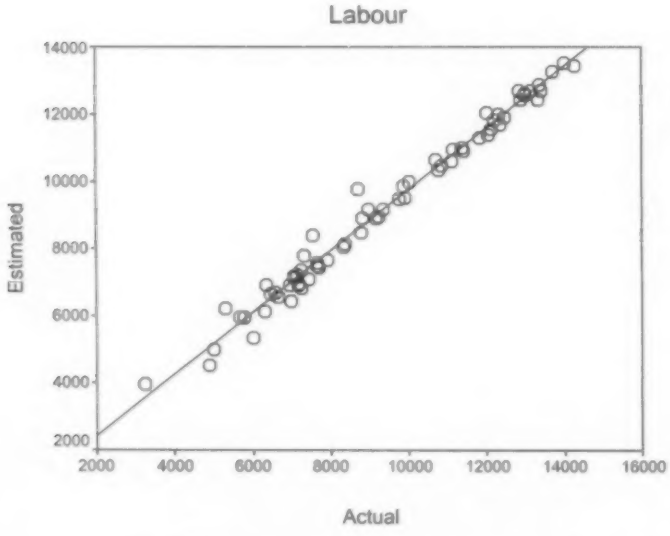

FIG. 3. Actual and Estimated Values of Labour Straight-ticket Voters, by Constituency.

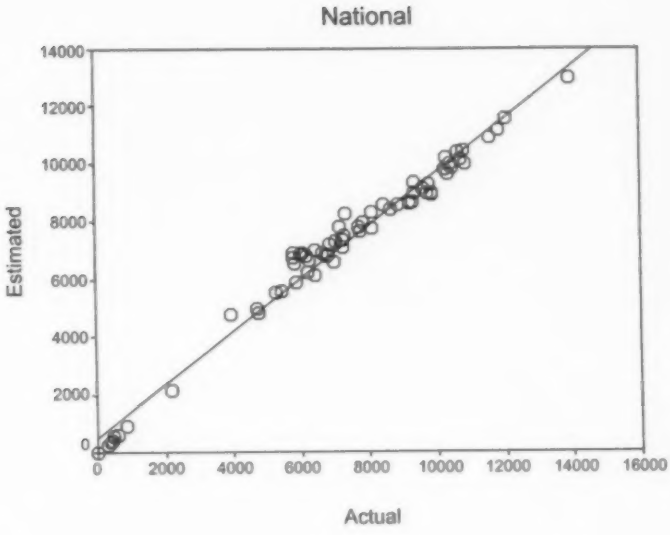

FIG. 4. Actual and Estimated Values of National Straight-ticket Voters, by Constituency.

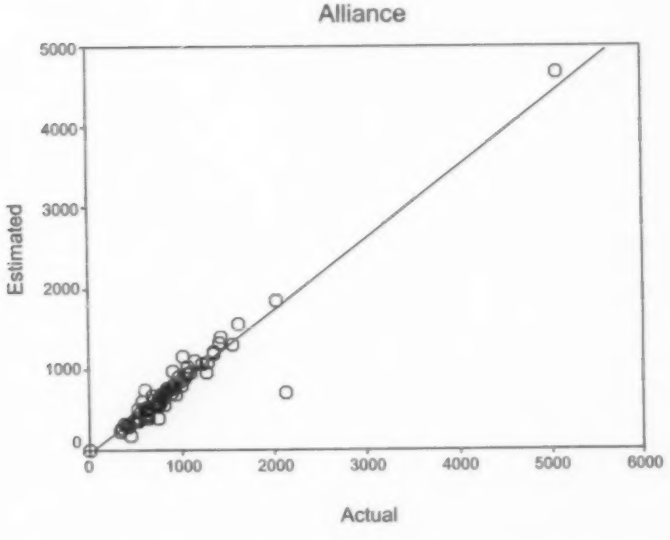

FIG. 5. Actual and Estimated Values of Alliance Straight-ticket Voters, by Constituency.

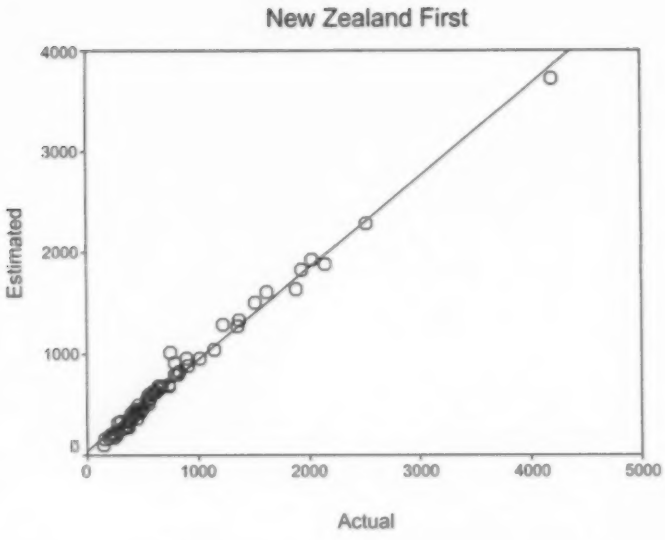

FIG. 6. Actual and Estimated Values of New Zealand First Straight-ticket Voters, by Constituency.

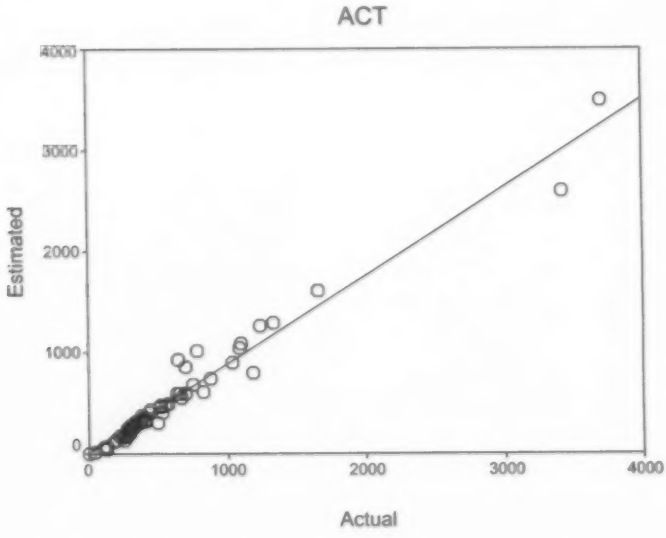

FIG. 7. Actual and Estimated Values of ACT Straight-ticket Voters, by Constituency.

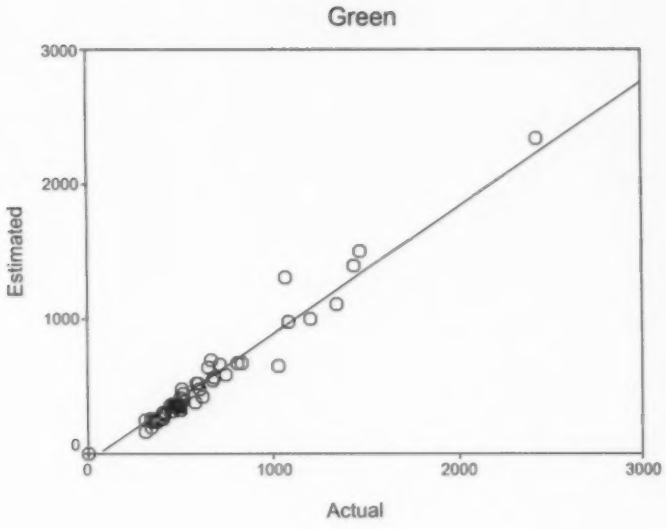

FIG. 8. Actual and Estimated Values of Green Straight-ticket Voters, by Constituency.

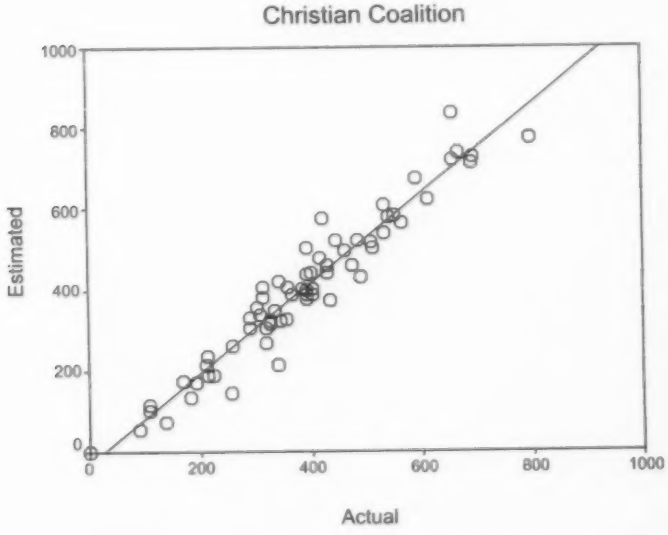

FIG. 9. Actual and Estimated Values of Christian Coalition Straight-ticket Voters, by Constituency.

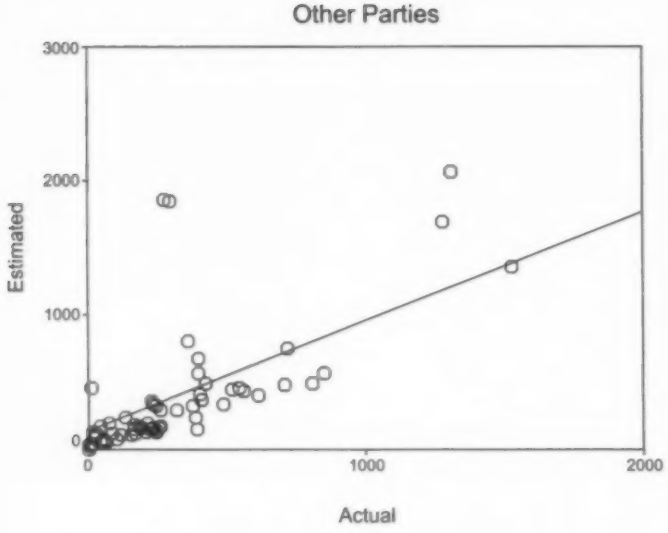

FIG. 10. Actual and Estimated Values of "Other" Straight-ticket Voters, by Constituency.

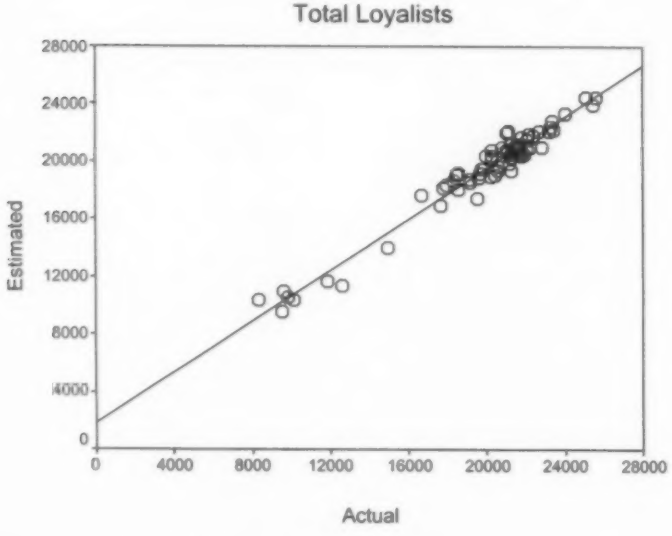

FIG. 11. Actual and Estimated Values of Total Number of Straight-ticket Voters, by Constituency.

fielded none). And electors who voted for one of those parties in the list contest may well have supported the candidate of a further "other" party in the constituency contest—preferring her/him instead one of the seven "major parties."

Overall, these results suggest that a great deal of confidence can be placed in the *EMax* estimates of the cells in the 67 constituency split-ticket matrices. No evidence is available to check the veracity of the off-diagonal estimates, but the circumstantial evidence from the diagonal values is very clear: they provide very good estimates of the number of electors who voted for one party in the list contest but for another in the constituency.

ACCOUNTING FOR THE GEOGRAPHY OF STRAIGHT- AND SPLIT-TICKETS

The purpose of making these estimates is not simply to map the variation in the volume of straight- and split-ticket voting across the constituencies but also to seek explanations for those variations. This section reports on preliminary steps leading towards such accounts, using first scatter diagrams to explore the nature of the expected relationships, and then regression analyses to test the veracity of the visual interpretations.

Visual Explorations

As noted above, split-ticket voting is likely to occur when the constituency candidate of the party voted for in the list contest has little chance of winning the seat, in which case the elector might switch her/his vote to the candidate of another party. In applying this general argument to the New Zealand situation in 1999, the relative situation of the various parties must be taken into account.

Before the change in the voting system in 1996, two parties—Labour and National—had dominated New Zealand politics for some sixty years. These continued to dominate the constituency contests at the first MMP election in 1996, with National winning 30 of the seats, and Labour 26: the Alliance, ACT, and United each won one seat (in each case their candidate was formerly prominent in one of the two main parties), and the populist New Zealand First Party won 6 (including the five Maori seats; their other was won by a former National Cabinet Minister). Apart from those that they won, the four "minor parties" represented in the House did not come very close to victory in any others: there was only one seat where ACT had a good chance of victory and none for the Christian Coalition; the Alliance had an outside chance (a 1996 vote of 20% or more) in eight and New Zealand First in six; the Green Party was part of the Alliance in 1996.

As general expectations, therefore, we expected to find: 1) that loyalty to each party (i.e., voting a straight ticket for each) would be greater in 1999, the greater the party's chances of victory in the constituency contest; and 2) that split-ticket voting from one of the "minor parties" to one of the "big two" would be larger, the greater the latter party's chances of victory in the constituency contest. To test whether this was so, we used the relevant party's percentage of the votes cast in 1996 in the constituency as a measure of its chances of victory there in 1999.[3]

Figures 12 and 13 are the scatter diagrams for straight-ticket (i.e., party loyalty) voting for each of the two largest parties. (In all of these scatter diagrams, we have fitted the lowest moving-average best-fit line, using 33% of the points.) In the case of Labour, the relationship is as expected: the better its performance in 1996, the larger

3. The constituency boundaries were changed between the two elections, and a further two constituencies created. Estimates of the result in 1996 as if the election then had been fought in the new constituencies were created by Alan McRobie using polling booth data: we are indebted to him for providing us with these data.

the percentage of its 1999 list voters who also supported the party's candidate in the constituency contest then (Figure 12). For the National party, the same relationship holds (the two constituencies where it fielded no candidate have been excluded) except that it retained the loyalty of a relatively low percentage of its list voters in most of its safest seats, producing an inverted-U relationship—although the downturn on the right-hand side of the graph is very much a function of a single observation only (Figure 13).

Turning to the "minor parties," there were two on the left of the political spectrum in 1999—the Alliance (dominated by former members of the Labour Party, who were stronger in their "socialist" leanings than those who had remained with Labour) and the Green Party (which was part of the Alliance in 1996 and fielded no constituency candidates then). Figure 14 shows, not surprisingly, that loyalty to the Alliance in 1999 was greatest in the only constituency that it would have won in 1996 (recall that we are using estimates of the 1996 results because of constituency changes: over 80% of those who voted for the Alliance list there also voted for the Alliance constituency candidate). Where the Alliance had little chance of victory, however, there was quite a wide spread of loyalty figures. The reason for this is shown in Figure 15, which indicates a clear positive relationship between the number of Alliance list voters who switched their constituency support to the Labour candidate and Labour's 1996 performance: the stronger Labour was in a constituency in 1996—and thus the better its chances of victory in 1999—the more support it picked up from those whose list vote went to the Alliance. If your party has no chance of success in the constituency, it seems, you vote for the party on the left of the political spectrum that does. And exactly the same applied to the Green Party list voters who switched to Labour in the constituency contest (Figure 16).

Turning to the two smaller parties on the right of the political spectrum—ACT (the Association of Consumers and Taxpayers) and the Christian Coalition—Figures 17 and 18 look at the percentage of their list voters who switched their constituency sup-

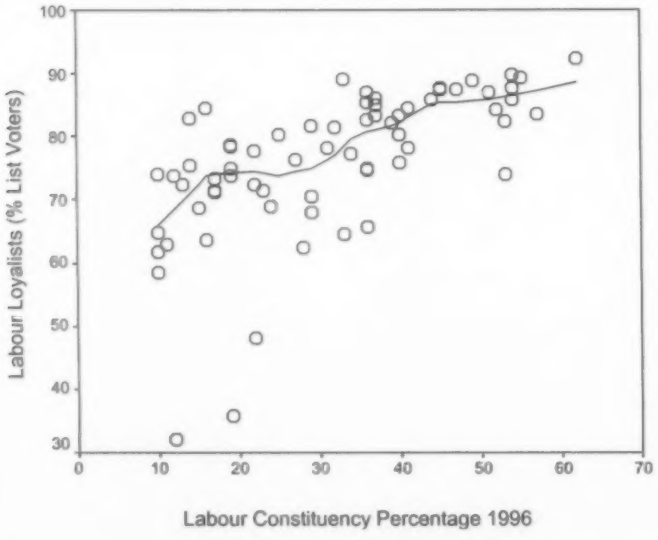

FIG. 12. Labour Constituency Vote Percentage in 1996 and Loyalists 1999, by Constituency.

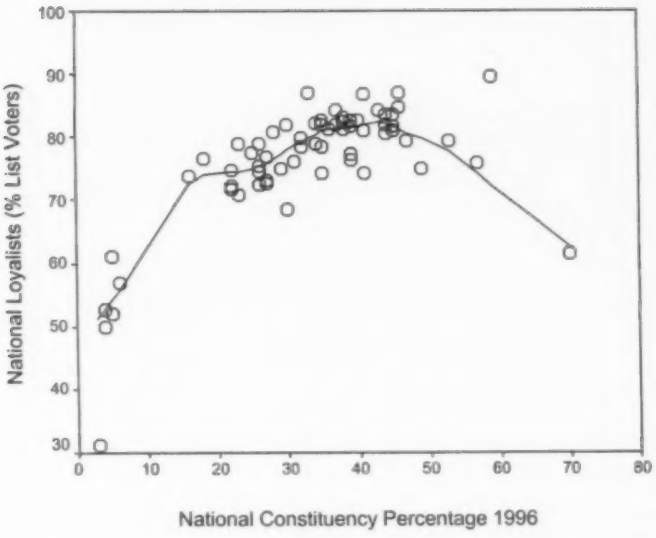

FIG. 13. National Constituency Vote Percentage in 1996 and Loyalists 1999, by Constituency.

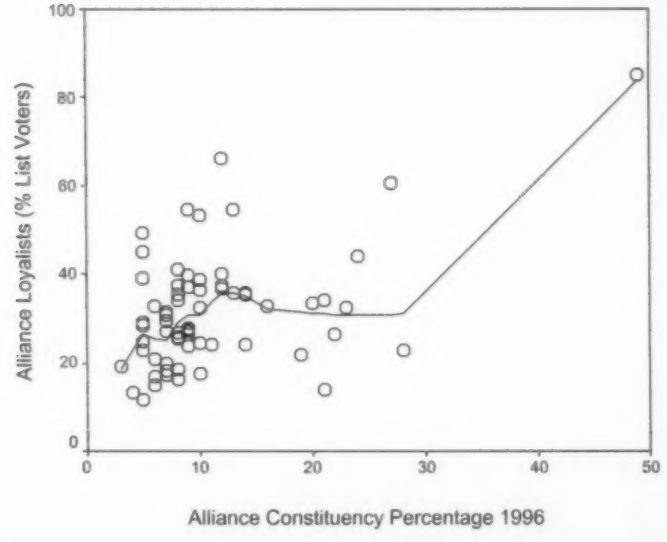

FIG. 14. Alliance Constituency Vote Percentage in 1996 and Loyalists 1999, by Constituency.

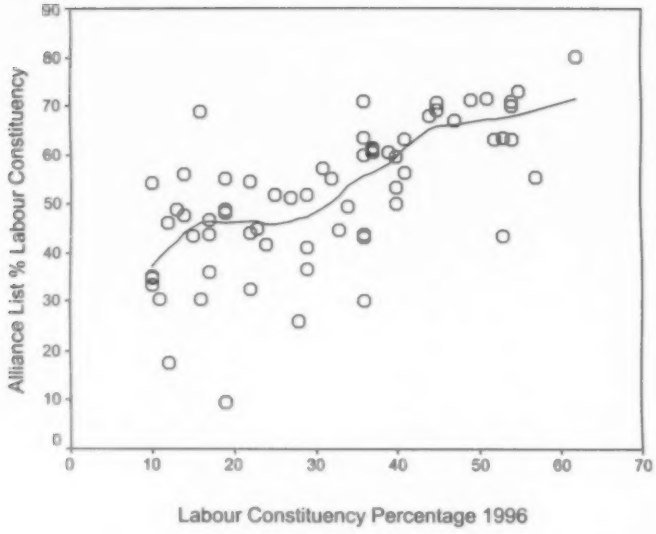

FIG. 15. Labour Constituency Vote Percentage in 1996 and Alliance-Labour Split-ticket Voters 1999, by Constituency.

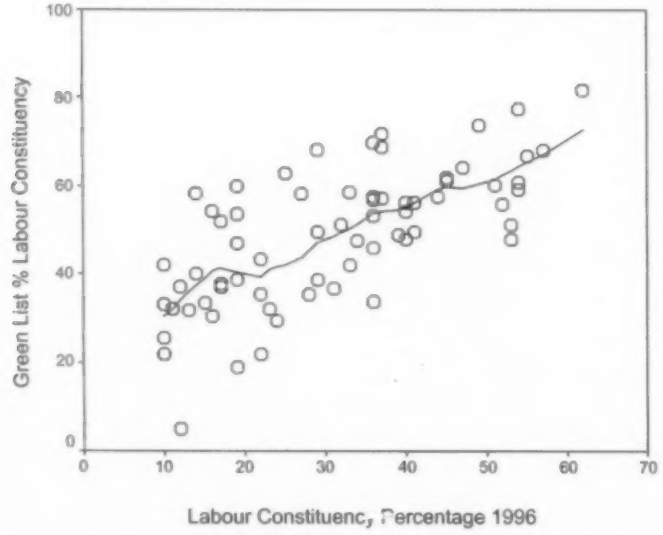

FIG. 16. Labour Constituency Vote Percentage in 1996 and Green-Labour Split-ticket Voters 1999, by Constituency.

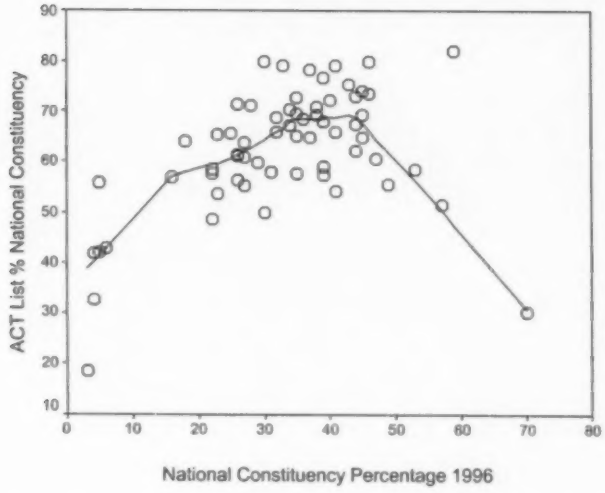

FIG. 17. National Constituency Vote Percentage in 1996 and ACT-National Split-ticket Voters 1999, by Constituency.

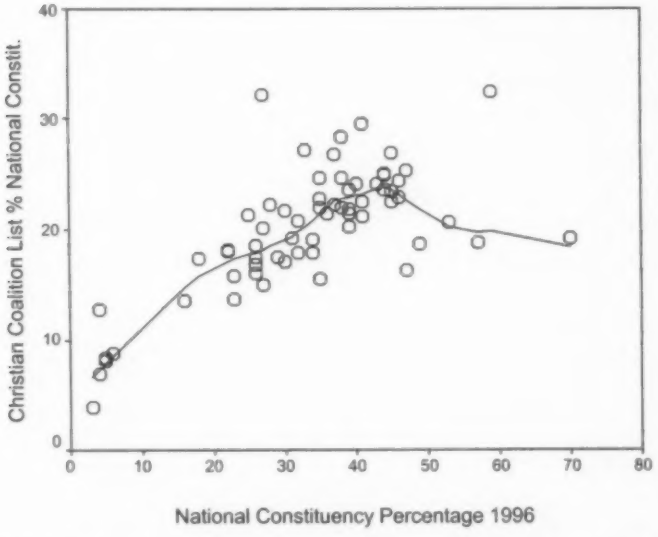

FIG. 18. National Constituency Vote Percentage in 1996 and Christian Coalition-National Split-ticket Voters 1999, by Constituency.

port to the right's major party—National. Again, the patterns are as expected—the better National's performance in 1996, the more who switched to it in 1999, although again National performed relatively poorly in its safest seat where it was subject to a strong challenge from the ACT candidate.[4]

Finally, there is the populist New Zealand First Party. This was initially formed as a breakaway party from National in the mid-1990s. After the 1996 election it held the balance of power between the two main parties and following prolonged negotiations agreed to enter a coalition with National. The coalition broke up in 1998, however, as did New Zealand First. In 1999, it won greatest loyalty from those who supported it in the list contest where its 1996 performance was best—although the scatter of points is fairly wide (Figure 19). Among those who voted a split ticket, however, Figures 20 and 21 suggest that most switched their support at the constituency contest to the Labour candidate, especially if he/she had a good chance of success: however, where the National candidate was well placed to win, there was a greater shift in that direction—although as Figure 20 shows, much of the positive relationship results from very small split-ticket shifts in that direction in the six constituencies where National performed especially badly in 1996: excluding those there is only a slight positive relationship, suggesting that after the 1998 coalition break-up most New Zealand First list supporters preferred Labour to National.[5] But the better Labour's chances of victory in a seat, the larger the percentage of New Zealand First's list voters who switched their support to the Labour candidate (Figure 21).

FIG. 19. New Zealand First Constituency Vote Percentage in 1996 and Loyalists 1999, by Constituency.

4. This was the Auckland constituency of Epsom, where the ACT candidate came a close second in 1999, having achieved only an estimated 3% of the votes in 1996. The ACT candidate in 1999 was an incumbent MP, who had won a list seat in 1996 (see Vowles 2000).

5. These are the six Maori seats, which would have been won by New Zealand First in 1996 but produced resounding victories for Labour in 1999 (Figure 20).

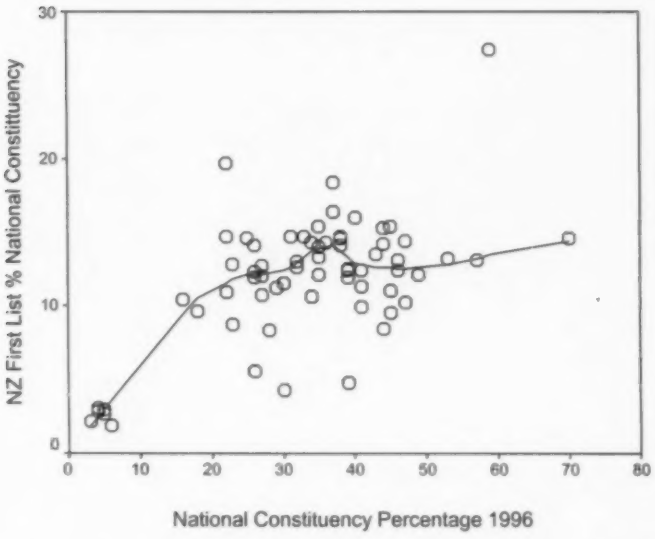

FIG. 20. National Constituency Vote Percentage in 1996 and New Zealand First-National Split-ticket Voters 1999, by Constituency.

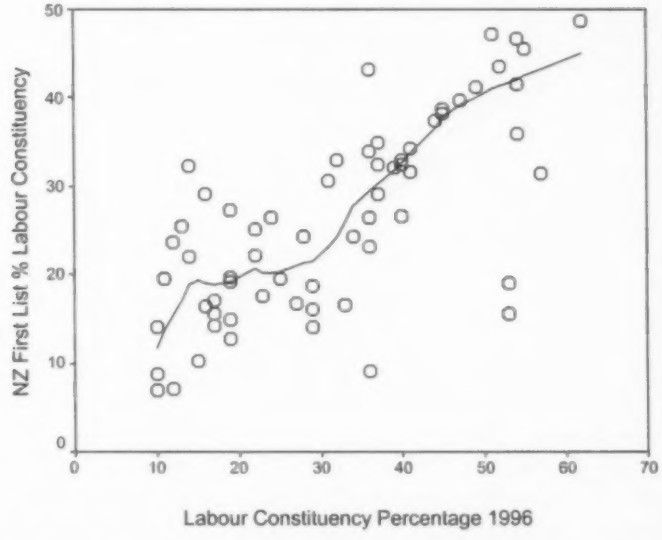

FIG. 21. Labour Constituency Vote Percentage in 1996 and New Zealand First-Labour Split-ticket Voters 1999, by Constituency.

Regression Analyses

These exploratory visual explorations suggest the validity of our hypotheses relating to the geography of straight- and split-ticket voting. This section reports on more formal tests, using OLS regression—and only fitting linear trends. In these, we test a further hypothesis, relating to campaign activity.

The logic underpinning the tests reported above is that voters act rationally in deciding whether to support their preferred party in their local constituency contest or instead vote a split ticket, switching their constituency support to another party. Some may well act in the latter way, having assessed their preferred party's chances of success in the constituency contest. But many will have been unprepared to take the time and effort to do that—to collect the relevant information (remember that the constituencies had changed since 1996). As theoreticians of democracy have been arguing since Downs (1957), voting choices have to be based on information, but the cost of collecting information may be considerable and many may feel it not worth the effort—given the low probability that how they vote will influence the result of the election. They are thus largely dependent on any information provided to them free, which takes no time to access and little to assess.

In constituency-based electoral systems, a major source of such information is the political parties and their candidates who are campaigning for votes. Their campaigns are designed to provide electors with the needed information, and to encourage them to support a particular candidate. A range of studies (including some on New Zealand: Johnston and Perry 1984; Johnston and Pattie 1999) have shown that in general the more that a party spends on its constituency campaign (mainly on leaflets and posters), relative to other parties, the better its performance there. At the 1999 general election, candidates in the constituency contest were limited to spending no more than NZ$20,000 on their campaigns, and we take these figures (provided by the Electoral Commission) as a measure of campaign intensity.

We ran the regressions in three stages:

1. Including just the constituency vote percentage in 1996 for the party supported in the list contest in 1999;
2. Adding the constituency vote percentage in 1996 for the party supported in the constituency contest in 1999; and
3. Adding the spending by the two parties involved at the 1999 constituency contest.

There is no second stage for the straight-ticket regressions. Our expectations were that the coefficients for the list party's 1996 vote would be positive, as would those for its own spending; the other two should be negative. For regressions involving the National Party these were run twice: once for all constituencies where it fielded a candidate in 1999 and the other excluding the aberrant case—Epsom—to ensure that this was not biasing any observed trend.

The results in Table 3 provide very strong support for the general model. The average r^2 value across the ten regressions at the third stage was 0.57; replacing the alternative regressions for National (in italics in Table 3) it is 0.61. In addition, all but two of the regression coefficients were significant at the 0.05 level or better (significant coefficients are shown in bold); the exceptions were those for ACT and Christian Coalition spending respectively in the regressions relating to splits from those to National. In general terms, therefore:

1. The better a party's chances of victory in a constituency, the larger the percentage of its list voters who also support its constituency candidate—i.e., vote a straight ticket for the party;
2. This tendency to vote a straight-ticket in constituencies that the party has a

TABLE 3

Regressions of Straight- and Split-ticket Votes on Party Performance in 1996 and Party Campaign Spending in 1999

Vote on List	Vote in Constituency	r^2 at stages 1	2	3	*b* coefficients (stage 3) for L66	C66	LSP	CSP
Labour	Labour	0.38	—	0.43	**0.45**	—	**0.67**	—
National	National	0.43	—	0.53	**0.34**	—	**0.73**	—
		0.61	—	*0.67*	***0.47***	—	***0.59***	—
Alliance	Alliance	0.22	—	0.51	**0.64**	—	**1.44**	—
Alliance	Labour	0.25	0.56	0.65	**−0.62**	**0.52**	**−0.88**	**0.56**
Green	Labour	—	0.54	0.64	—	**0.57**	**0.61**	**−1.23**
ACT	National	0.15	0.19	0.32	**0.23**	**−1.26**	0.17	**1.14**
		0.03	*0.43*	*0.54*	***−1.16***	**0.49**	*−0.33*	***0.90***
Christian Coaln.	National	0.49	0.50	0.62	**−0.58**	**0.20**	−0.13	**0.48**
		0.00	*0.56*	*0.65*	***−0.59***	***0.24***	*−0.09*	***0.44***
NZ First	NZ First	0.36	—	0.66	**0.22**	—	**2.08**	—
NZ First	Labour	0.29	0.60	0.66	**−0.15**	**0.41**	**−0.70**	**0.34**
NZ First	National	0.36	0.60	0.70	**−0.09**	**0.11**	**−0.33**	**0.16**
		0.50	*0.61*	*0.71*	***−0.09***	***0.12***	***−0.36***	***0.14***

NOTES: Key to variables: L66—list party's constituency vote percentage in 1966; C66—constituency party's constituency vote percentage in 1966; LSP—list party's constituency spending in 1999; CSP—constituency party's constituency spending in 1999. Significant regression coefficient sat the 0.05 level are shown in bold. The regressions in italics are those involving the National party and omitting the constituency of Epsom.

good chance of winning is accentuated in those where it also spends heavily on its local campaign—the greater the intensity of that campaign, the greater the number of straight-tickets as a percentage of its list supporters;

3. The better one of the two major parties' chances of success in the constituency contest, the larger the percentage of those who supported another party on its side of the political spectrum who switched to voting for it in the constituency contest—i.e., split their ticket in favor of the stronger party on that side of the spectrum;
4. This ability to attract split tickets in constituencies where they had good chances of success was accentuated in those constituencies where the major parties campaigned most intensively;
5. The success of the large parties' campaigns in attracting split-ticket votes was mitigated to some extent by the intensity of the smaller parties' local campaigns—the more that the small parties spent, the smaller the outflow of split-ticket votes to the larger party (whatever its spending), although this conclusion applied only to the parties of the left (Alliance and Green)

The geography of split- and straight-ticket voting in New Zealand in 1999 was consistent witch our hypotheses.

CONCLUSION

Most attention with regard to the ecological inference problem has focused on what we have termed the empty cell situation (Figure 1), where there is no prior knowledge of the internal cell values of any matrix within a data cube. Estimating those unknown values is the goal—with special attention focused on the unknown values in the global matrix, although the values in the individual area matrices are of considerable interest to geographers and others concerned with spatial variations in the behavior patterns under consideration.

There are, however, other ecological inference problems, where the overall focus is the same—estimating unknown values in the internal cells of a data cube—but further information is available on which the estimates can be based, which do not exist

for the "classic" empty cell situation. One of these is what we have termed the sum-of-cells situation (Figure 2), in which a further set of information is available at the global scale. This can be used as a third constraint in procedures designed to solve the ecological inference problem, and thereby produce estimates of spatial variation in behavior which are consistent with the overall pattern in that behavior.

A method for resolving the ecological inference problem in this sum-of-cells situation (*EMax*) has been in use for some time, but it has not been possible to evaluate the accuracy of its estimates of the geography of behavior in any major way. This difficulty has been overcome in the analyses reported here, using data on straight- and split-ticket voting at the 1999 New Zealand general election. The estimates of straight-ticket voting produced by *EMax*, given the availability of data for the three sets of constraints, have been compared with the "real" values collated by the Electoral Commission, and found to be extremely accurate. For this data set at least, we have great confidence in the estimated off-diagonal, split-ticket voting estimates, given the obvious success of the procedure in estimating the on-diagonal, straight-ticket votes.

The strength of these findings suggests that there are clear benefits to be gained from wider application of the *EMax* methodology, including comparisons of its performance—where relevant—with that of *EI*. There is a wide range of potential applications for both of these procedures in geography (as Sui et al. 2000, identified for *EI*), exploration of which might lead to ways of merging the two.

The purpose of solving the ecological inference problem in geographical studies is to enable studies to be undertaken of spatial variations in the behavior under consideration, and to test hypotheses regarding the patterning of that variation. That has been illustrated here with analyses of the geography of straight- and split-ticket voting at New Zealand's 1999 general election. We have been able to show that those patterns are consistent with expectations based on rational choice models. People who support a party in the list contest are more likely to support its candidate in the constituency contest as well if he/she has a good chance of success there—and are particularly likely to do so if that candidate campaigns hard for their support.

LITERATURE CITED

Berg, S. (1988). Spatial Influences on Voter Transitions in Sweden: An Application of Johnston's Maximum Entropy Method. *Electoral Studies* 7, 233–50.

Burden, B. C., and D. C. Kimball (1998). A New Approach to the Study of Ticket-Splitting. *American Political Science Review* 92, 533–44.

Cho, W. T. (1998). If the Assumptions Fit. *Political Analysis* 7, 143–63.

Cleave, N., P. J. Brown, and C. Payne (1995). Evaluation of Methods for Ecological Inference. *Journal of the Royal Statistical Society, Series A* 158, 55–72.

Deming, W. E., and F. F. Stephan (1940). On a Least Squares Adjustment of a Sampled Frequency Table When Expected Marginal Totals Are Known. *Annals of Mathematical Statistics* 11, 427–44.

Downs, A. (1957). *An Economic Theory of Democracy*. New York: Harper and Row.

Freedman, D. A., S. P. Klein, and D. Ostland (1998). Review of *A Solution to the Problem of Ecological Inference*. *Journal of the American Statistical Association* 93, 1518–22.

Gelman, A., D. K. Park, S. Ansalobehere, P. N. Price, and L. C. Minnite (2001). Models, Assumptions and Model Checking in Ecological Regression. *Journal of the Royal Statistical Society, Series A* 164, 101–18.

Goodman, L. A. (1959). Some Alternatives to Ecological Correlation. *American Journal of Sociology* 64, 610–24.

Johnston, R. J., and A. M. Hay (1982). On the Parameters of Uniform Swing in Single-Member Constituency Electoral Systems. *Environment and Planning A* 14, 61–74.

Johnston, R. J., and A. M. Hay (1983). Voter Transition Probability Estimates: An Entropy-Maximizing Approach. *European Journal of Political Research* 11, 93–98.

Johnston, R. J., A.M. Hay, and D. Rumley (1983). Entropy-Maximizing Method for Estimating Voting Data: A Critical Test. *Area* 15, 35–41.

Johnston, R. J., A. M. Hay, and D. Rumley (1984). On Testing for Structural Effects in Electoral Geography, Using Entropy-Maximizing. *Environment and Planning A* 16, 233–40.

Johnston, R. J., and C. J. Pattie (1999). Constituency Campaign Intensity and Split-Ticket Voting: New Zealand's First Election under MMP, 1996. *Political Science* 51, 164–81.

Johnston, R. J., and C. J. Pattie (2000). Ecological inference and entropy-maximizing: an alternative procedure for split-ticket voting. *Political Analysis* 8, 333–45.

Johnston, R. J., and C. J. Pattie (2001). On Geographers and Ecological Inference. *Annals of the Association of American Geographers* 91, 281–82.

Johnston, R. J., C. J. Pattie, and J. G. Allsopp (1988). *A Nation Dividing? Britain's Changing Electoral Map, 1979–1987*. London: Longman.

Johnston, R. J., and P. J. Perry (1984). Campaign Spending and Voting in the New Zealand General Election 1981: A Note. *New Zealand Geographer* 39, 81–83.

King, G. (1997). *A Solution to the Ecological Inference Problem: Recovering Individual Behavior from Aggregate Data*. Princeton, N.J.: Princeton University Press.

King, G. (1999). The Future of Ecological Inference Research. *Journal of the American Statistical Association* 94, 352–55.

King, G., O. Rosen, and M. A. Tanner (1999). Binomial-Beta Hierarchical Models for Ecological Inference. *Sociological Methods & Research* 28, 61–90.

Kousser, J. M. (2001). Ecological Inference from Goodman to King. *Historical Methods* 34, 101–26.

Lewis, J. B. (2001). Understanding King's Ecological Inference Model: A Method-of-Moments Approach. *Historical Methods* 34, 170–88.

Mosteller, F. (1968). Association and Estimation in Contingency Tables. *Journal of the American Statistical Association* 63, 1–28.

O'Loughlin, J. (2000). Can King's Ecological Inference Method Answer a Social Scientific Puzzle: Who Voted for the Nazi Party in Weimar Germany? *Annals of the Association of American Geographers* 90, 592–600.

Robinson, W. S. (1950). Ecological Correlations and the Behavior of Individuals. *American Sociological Review* 15, 351–57.

Rosen, O., W. Jiang, G. King, and M. A. Tanner (2001). Bayesian and Frequentist Inference for Ecological Inference: The *R* x *C* Case. *Statistica Neerlandica* 55, 134–56.

Sui, D. L., A. S. Fotheringham, L. Anselin, J. O'Loughlin, and G. King (2000). Book Review Forum. *Annals of the Association of American Geographers* 90, 579–606.

Vowles, J. (2000). The Impact of the 1999 Campaign. In *Left Turn: the New Zealand General Election of 1999*, edited by J. Boston et al., 141–60. Wellington: Victoria University Press.

Wilson, A. G. (1967). A Statistical Theory of Spatial Distribution Models. *Transportation Research* 1, 253–69.

Wilson, A. G. (1970). *Entropy in Urban and Regional Modelling*. London: Pion.

Theo Arentze
and Harry Timmermans

Measuring Impacts of Condition Variables in Rule-Based Models of Space-Time Choice Behavior: Method and Empirical Illustration

The use of rule-based systems for modeling space-time choice has gained increasing research interests over the last years. The potential advantage of the rule-based approach is that it can handle interactions between a large set of predictors. Decision tree induction methods are available and have been explored for deriving rules from data. However, the complexity of the structures that are generated by such knowledge discovery methods hampers an interpretation of the rule-set in behavioral terms with as a consequence that the models typically remain a black box. To solve this problem, this paper develops a method for measuring the size and direction of the impact of condition variables on the choice variable as predicted by the model. The paper illustrates the method based on location and transport-mode choice models that are part of Albatross *model—an activity-based model of space-time choice.*

1. INTRODUCTION

The recent planning literature has witnessed the exploration of decision tree techniques for analyzing space-time behavior. Decision tree induction methods, which have been developed in statistics and artificial intelligence, provide a technique to derive decision trees from observations on choice. The application of the technique for developing predictive models of space-time choice has been explored in several recent studies (Thill and Wheeler 1999; Gahegan 2000a; Arentze et al. 2000; Wets et al. 2000; Yamamoto, Kitamura, and Fujii 2002; Moons et al. 2002). In particular, we have used the technique to develop the activity-based model Albatross (Arentze and Timmermans 2000, 2002). Several empirical studies showed that the technique for such purposes is promising in terms of internal validity, performance relative to conventional utility-based models and spatial transferability (Arentze et al. 2001). Gahegan

Harry Timmermans is professor in the Department of Architecture and Urban Planning at Eindhoven University of Technology. E-mail: h.j.p.timmermans@bwk.tue.nl. *Theo Arentze is an assistant professor in the Department of Architecture and Urban Planning at Eindhoven University of Technology. E-mail:* t.a.arentze@bwk.tue.nl.

Geographical Analysis, Vol. 35, No. 1 (January 2003) The Ohio State University
Submitted: February 21, 2002. Revised version accepted: June 27, 2002.

(2000b) stresses the potential of inductive learning as a method, besides scientific visualization, to address data with high attribute dimensionality.

The strength of the decision tree technique for modeling space-time behavior is its ability to represent the full complexity of interactions between condition variables. Unlike algebraic models (e.g., Bowman and Ben-Akiva, 1999), the specification of the model does typically not rely on the commonly used linear-additive utility functions, possibly with some selected interaction effects. At the same time, however, this complexity hinders an interpretation in terms of the impacts of condition variables on the choice variable. Whereas it is relatively straightforward to derive elasticities in discrete choice models, the non-algebraic nature of decision-tree based models prevents one from such easy interpretation.

Measuring the importance of condition variables has received some attention in Statistics and Artificial Intelligence mainly for two applications. First, it has been considered for identifying the most significant features among a large set of potentially relevant explanatory variables in a pre-modeling stage (e.g., Kira and Rendall 1992; Kononenko 1994). Feature selection methods that have been proposed take into account possible correlations between condition variables and try to measure the independent contribution of condition variables to observed differences in choice for example by using information theoretical concepts (e.g., Piramuthu 1996). In a second application, the objective is to rank conditional variables according to the contribution that they make in classifying or predicting the target variable in a decision tree model (e.g., Steinberg and Colla 1997). Thus, in both applications attention is focused on the potential or apparent significance of variables for prediction or classification. We emphasize that methods that have emerged from this area are not transferable to the present problem. Elasticity refers to a different dimension, namely the extent to which predictions are sensitive to variation over a hypothetical range of the condition variable of interest. To the best of our knowledge, methods for deriving such information from rule-based models do not exist.

To fill this gap, the aim of this paper is to introduce a method to derive elasticity information from rule-based models. The principle of the proposed method is straightforward. After having derived a rule-based model from training data, the model is used to predict a choice frequency distribution multiple times for a same set of cases under different simulated settings of one condition variable at a time. Then, elasticity or impact is conceptualized as the size of the difference in choice frequency distribution between levels of the condition variable. Because a model, rather than data, and simulated variation over the full range of a condition variable, rather than existing variation, is used the resulting measures can be interpreted as elasticity or impact coefficients.

The proposed method is illustrated using decision tree models for location and transport mode choice that are part of the Albatross model system. The original version of Albatross was derived from a limited study area. The results presented here are based on a re-induction of the system conducted to extend the study area to the national scale. Given the large scale of the data set used, the elasticities derived from the models should provide potentially useful information for time geographical research.

To introduce and illustrate the method, the paper is structured as follows. The next section first describes the format of rule-based models as assumed in the method. Then, the second section introduces the proposed method for extracting elasticity coefficients. The section that follows discusses the results of the case study to illustrate the method. Finally, the paper concludes with summarizing major findings.

2. A REPRESENTATION FORMAT OF RULE-BASED MODELS

The method developed in the next section is suitable for rule-based models based on the decision tree formalism, but is not restricted to that formalism. Therefore, this section describes requirements for a rule-based formalism in a more generic way.

The assumed model considers a discrete set of choice alternatives i and predicts a choice probability for each alternative i. The model consists of a set of decision rules possibly (but not necessarily) structured as a decision tree and predicts a choice probability for each given case as:

$$p_i(\mathbf{c}) = \sum_k \Pr(k|\mathbf{c}) \Pr(i|k) \qquad \forall i \tag{1}$$

where k is an index of a decision rule, c is a vector of condition values observed in the given case, $\Pr(k \mid \mathbf{c})$ is the probability that c meets the condition state of rule k, and $\Pr(i \mid k)$ is the probability that rule k generates prediction i. To be consistent and complete, the rule-set should meet the following conditions:

$$\sum_k \Pr(k|\mathbf{c}) = 1 \qquad \forall \mathbf{c} \tag{2}$$

$$\sum_i \Pr(i|k) = 1 \qquad \forall k \tag{3}$$

In conventional rule-based formalisms, which we consider here, a rule is activated in a deterministic way implying that $\Pr(k \mid \mathbf{c})$ is either zero or one in equations (1) and (2). The condition state of each rule defines a conjunct set of necessary conditions. So, the rule selection function can be written as:

$$\Pr(k|\mathbf{c}) = \begin{cases} 1, \text{ if } \; c_1 \in CS_{1k} \wedge c_2 \in CS_{2k} \wedge \ldots \wedge c_n \in CS_{nk} \\ 0, \text{ otherwise} \end{cases} \tag{4}$$

where $CS_{sk} \subseteq CD_s$ is the state of the s-th condition variable in the k-th rule, CD_s is the domain of the s-th condition variable and $\mathbf{c} = (c_1 \ldots c_n)$. If the condition variable is continuous, a state is generally defined as a certain subrange of CD_{sk}. Otherwise, if the condition variable is discrete, a distinction is made between ordinal and nominal variables. A state of an ordinal variable is restricted to subsets of the domain that preserve the order of the elements, whereas a state of a nominal variable can take on any subset of the domain. Format (4) does not necessarily imply that every condition variable is relevant in every rule. Irrelevance of a variable is represented by a special definition of its state. That is, if s is irrelevant in the k-th rule we define $CS_{sk} = CD_s$. This means that every possible value of c_s meets the condition so that in fact c_s is irrelevant for determining whether the case matches k.

Decision tree induction is one of several possible methods for finding structures of type (4) that best fit data on observed choice. The term 'training set' is generally used to refer to the sample of observations used to derive a rule-set. The fit of a decision tree is conceptualized as the degree of homogeneity or purity of choice outcomes at leaf nodes. Several heuristic methods have been proposed in AI and statistics to find the optimal tree. Top-down methods develop a tree by recursively splitting the training set on condition variables until a single choice outcome is observed within partitions or no further improvement is possible. Different criteria of splitting have been proposed including a measure of information gain (C4.5, Quinlan 1993), the Gini-index (CART, Breiman et al. 1984) and a Chi-square based measure of significance of difference between partitions (CHAID, Kass 1980). Other differences between

methods exist, but in the context of this study the process by which trees, or rule sets in general, are derived is not relevant as long as the format confirms to definitions (1)–(4).

The second term on the right hand side of equation (1) defines the predicted choice probability of i for any case matching rule k. In decision tree applications, an all-or-nothing action assignment rule is commonly used. The so-called 'plurality' rule assigns the modal choice of training cases at leaf node k to each unseen case assigned to k. Such a rule does maximize expected probability of correct predictions, but fails to reproduce remaining variance in choice outcomes, if any, at leaf nodes. In planning applications, it is generally important to reproduce unexplained variance in predictions. To come up to that need, Albatross uses a probabilistic assignment rule instead. Hereby, choice probabilities are derived from the training set as:

$$\Pr(i|k) = \frac{f_{ki}}{n_k} \tag{5}$$

where f_{ki} is the frequency of choice i at leaf node k and n_k is the total number of cases at leaf node k in the training data. However, in the present context, it is not important how the action assignment rule is derived as long as constraint (3) is met.

3. DERIVING IMPACT TABLES

Given a decision tree of the form as defined by (1)–(4), this section outlines a method of deriving impact measures for each condition variable s. We emphasize that the purpose of the measures is not to replace the rule-based model, but rather to facilitate an interpretation of the model in behavioral terms. The method relies on the principle of simulating changes in a representative sample of cases on one condition variable at a time and next evaluating the impact on choice predictions. The sample should be representative for the population to which the model is applied. In many cases, the training set can be used for this. In this section we describe the method in more detail.

Following our earlier notation, the sample (e.g., training data) can be represented as:

$$f_{u(s),i} = \sum_j p_i(\mathbf{c}_{u(s),j}) \qquad \mathbf{c}_j = (c_{1j}, c_{2j}, \ldots, c_{nj}) \qquad \forall j \tag{6}$$

where as before s is an index of condition variables and j is an index of cases. The method assumes that each condition variable is a discrete variable with levels $c_{sj} = 1 \ldots m(s)$. This assumption does not exclude continuous variables from the analysis, but rather implies that these variables should be discretisized in a pre-processing step. For example, one can use an equal-frequency method, which divides a continuous scale into m intervals in such a way that each interval contains an equal number of cases. If q is the number of choice alternatives, the model is used to predict an $m(s) \times q$ frequency table for each condition variable s as:

$$i = 1 \ldots q, \text{ for } u(s) = 1 \ldots m(s)\ , \tag{7}$$

where $\mathbf{c}$ is a manipulated vector of condition values and p_i is the predicted choice probability of i using equation (1). The condition vector in (7) is defined as:

$$\mathbf{c}_{u(s)j} = (c_{1j}, c_{2j}, \ldots, u(s), \ldots c_{nj}) \qquad \forall j \tag{8}$$

This means that for deriving the impact frequency table a prediction run is repeated $m(s)$ times replacing for each case the observed value of the s-th condition variable by a varying value $u(s) = 1....m(s)$ and keeping all other variables as observed.

The frequency table resulting from this procedure contains the full information on the impact of condition variable s on choice outcomes. We propose to derive the following impact measures from the table:

$$IS_s = D(\boldsymbol{F}_s, \overline{\boldsymbol{F}}_s) \tag{9}$$

$$IS_{si} = D(\boldsymbol{F}_{si}, \overline{\boldsymbol{F}}_{si}) \tag{10}$$

$$MS_{si} = \frac{\sum_{u(s)=2}(f_{u(s),i} - f_{u(s)-1,i})}{\sum_{u(s)=2}|f_{u(s),i} - f_{u(s)-1,i}|} \tag{11}$$

where $D(\bullet,\bullet)$ is some measure of difference between frequency tables, $\boldsymbol{F}_s$ is the $m(s) \times q$ frequency table predicted by the model, $\overline{\boldsymbol{F}}_s$ is the expected frequency table under the assumption of no impact of s, $\boldsymbol{F}_{si}$ is the i-th column of the frequency table generated by the model, $\overline{\boldsymbol{F}}_{si}$ is the expected counterpart under the assumption of no impact of s on choice i, and $f_{u(s),i}$ is the $u(s) \times i$ cell frequency of the table predicted by the model. Several goodness-of-fit measures can be specified for D such as for example Chi-square, the sum of absolute differences, likelihood and so on.

As implied by the equations, the first measure gives an indication of the overall impact of variable s on the choice variable and the second measure provides a disaggregation to obtain an indication of its impact on each choice alternative i. Finally, the third measure is included to obtain information on the functional shape of the measured impact. More specifically, the measure can be interpreted as a measure of monotonicity. To see this, consider the following special cases. If the condition variable has a monotonically increasing impact on choice frequency across the levels of the condition variable, then MS_{is} equals 1, because in that case:

$$\sum_{u(s)=2}|f_{u(s),i} - f_{u(s)-1,i}| = \sum_{u(s)=2}(f_{u(s),i} - f_{u(s)-1,i}) \tag{12}$$

whereas, if the condition variable has a monotonically decreasing impact, the ratio equals -1, because in that case:

$$\begin{aligned}\sum_{u(s)=2}|f_{u(s),i} - f_{u(s)-1,i}| &= \sum_{u(s)=2}(-f_{u(s),i} + f_{u(s)-1,i}) \\ &= -\sum_{u(s)=2}(f_{u(s),i} - f_{u(s)-1,i})\end{aligned} \tag{13}$$

In all other cases, the ratio has a value in between -1 and 1. To see how the ratio behaves within this range, it is clarifying to rewrite (12) as:

$$\sum_{u(2)=2}(f_{u(s),i} - f_{u(s)-1,i}) = \sum_{u(s)=2} f_{u(s),i} - \sum_{u(s)=2} f_{u(s)-1,i} = -f_{1,i} + f_{m(s),i} \tag{14}$$

This means that MS_{si} is sensitive only to the difference in choice frequency between the first and last level of the condition variable. If the impact function has a perfectly U-shape or reversed U-shape, MS_{si} is zero. In all other cases, the measure represents some degree of negative or positive monotonicity. This has a clear meaning for condi-

tion variables that are measured on an ordinal or higher level or for binary condition variables. However, in the case of multi-level nominal variables, the order of levels is purely arbitrary so that the meaning of the measure is limited.

Measures (9), (10), and (11) are complementary and will be used to derive substantial information from a rule-based model. Statistical significance tests are considered not relevant in this application, as there are no error terms associated with the measures. The procedure is rather computationally demanding but still can be completed in reasonable time. For a given rule-set, it requires repeating a prediction run $m(s)$ times for each condition variable s.

4. ILLUSTRATION

This section illustrates the derivation and use of impact tables with some representative decision trees concerning location and mode choice that are part of Albatross, a rule-based model of travel demand, originally developed for the Dutch Ministry of Transport, which predicts which activities will be conducted where, when, for how long, with whom, and the transport mode involved (Arentze and Timmermans 2000, 2002). It is one of the most complex activity-based models that currently exists (see Timmermans, Arentze, and Joh 2002 for a review). It is not the aim of this paper to discuss the model in detail, but we do need to provide some background information to fully appreciate the context of this illustration. We therefore, first briefly discuss the process model in Albatross, to provide a context for discussion of results in the sections that follow.

4.1. The Albatross *process model*

Albatross consists of two major components that together define an activity schedule for each individual and each day. The first component generates an activity skeleton consisting of activities that are supposedly fixed in the short-term, their exact start time, and duration. Given the skeleton, the second component then determines the part of the schedule related to flexible activities to be conducted that day, their travel party, duration, time-of-day, and travel characteristics. Both components make use of the same location model component determining the location of activities. All three components assume a sequential decision process in which key choices are made and pre-defined rules operate to delineate choice sets and implement made choices in the current schedule.

Figures 1–3 schematically present the structure of each of the main components of the model. Each numbered rectangle corresponds to a decision tree, which is derived from activity diary data. The indices used in the figures are defined as:

i	index of activity in order of priority, $i = 1 \ldots I$
j	index of episode of activity i in order of start time, $j = 1 \ldots J$
k	index of tour in order of start time, $k = 1 \ldots K$

The skeleton component comprises decisions 1–13, the location component used for both fixed and flexible activities decisions 14–20 and the flexible activity component decisions 21–27.

Clearly, there are many ways in which the scheduling problem can be broken down into key choices and key choices can be arranged into a sequential decision process. Two principles have guided the design of the process model. First, qualitative decisions are made as much as possible explicitly. For that reason, decisions such as the selection of an activity, the number of episodes per activity, trip linkages between activity episodes, and relative locations of activity episodes are made separately prior to determining their exact attributes in terms of start time, duration, location, and

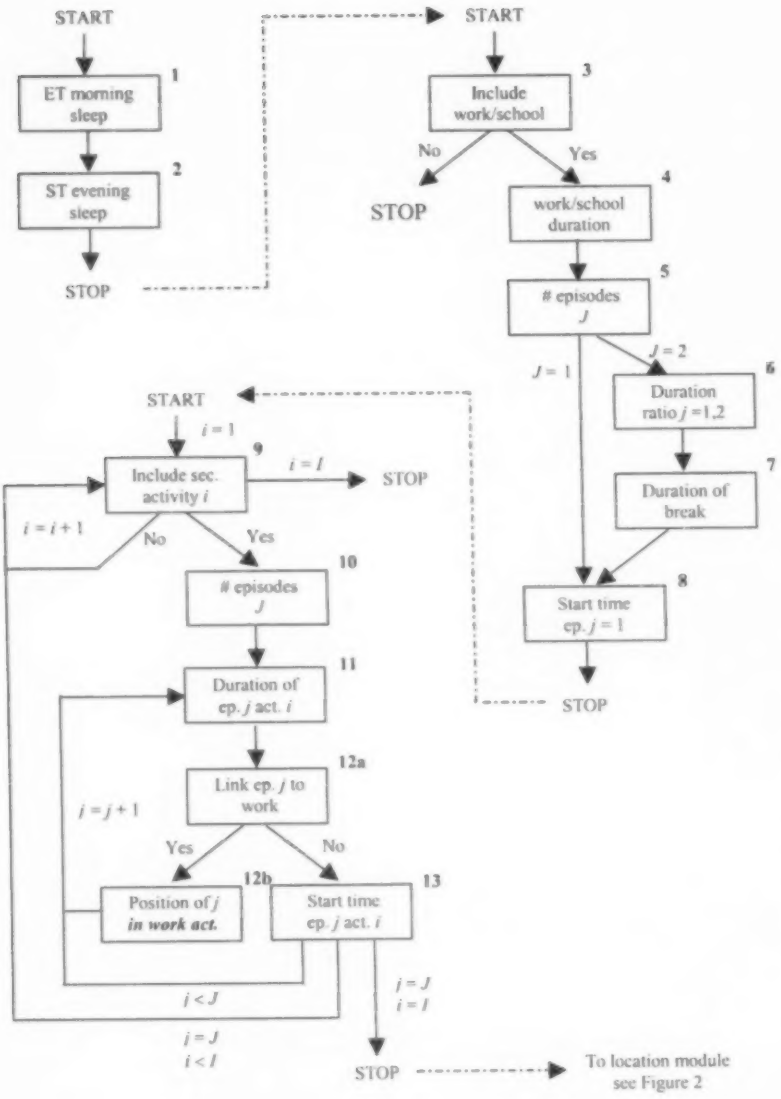

FIG. 1. The Process Model for Generating Fixed Activity Patterns

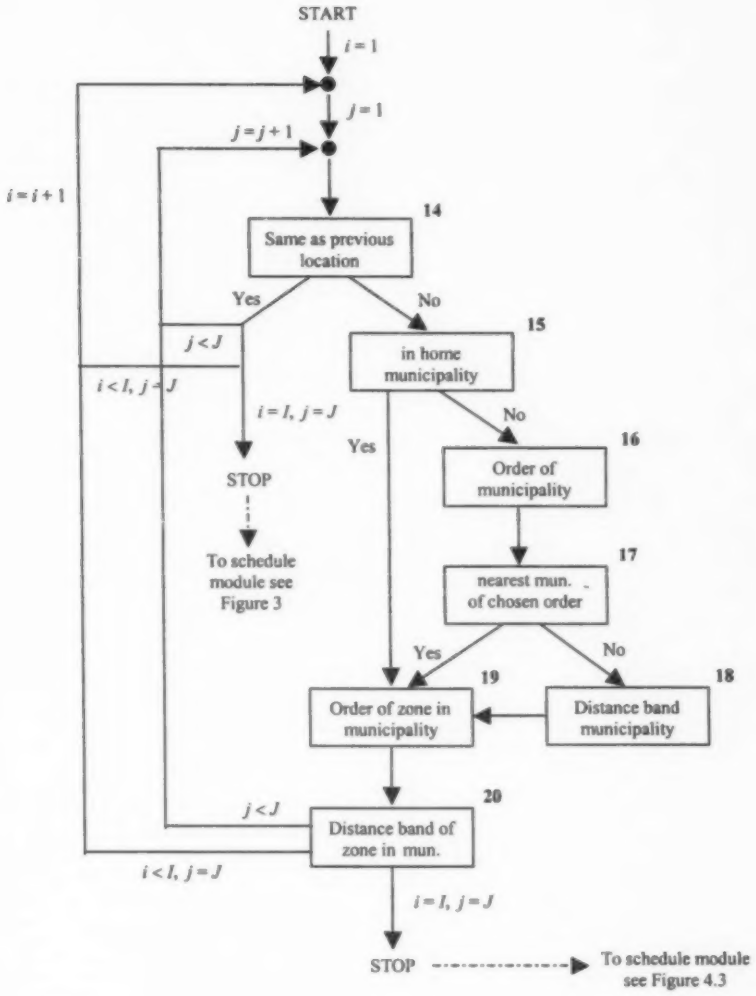

FIG. 2. The Process Model for Predicting Locations of Fixed and Flexible Activities

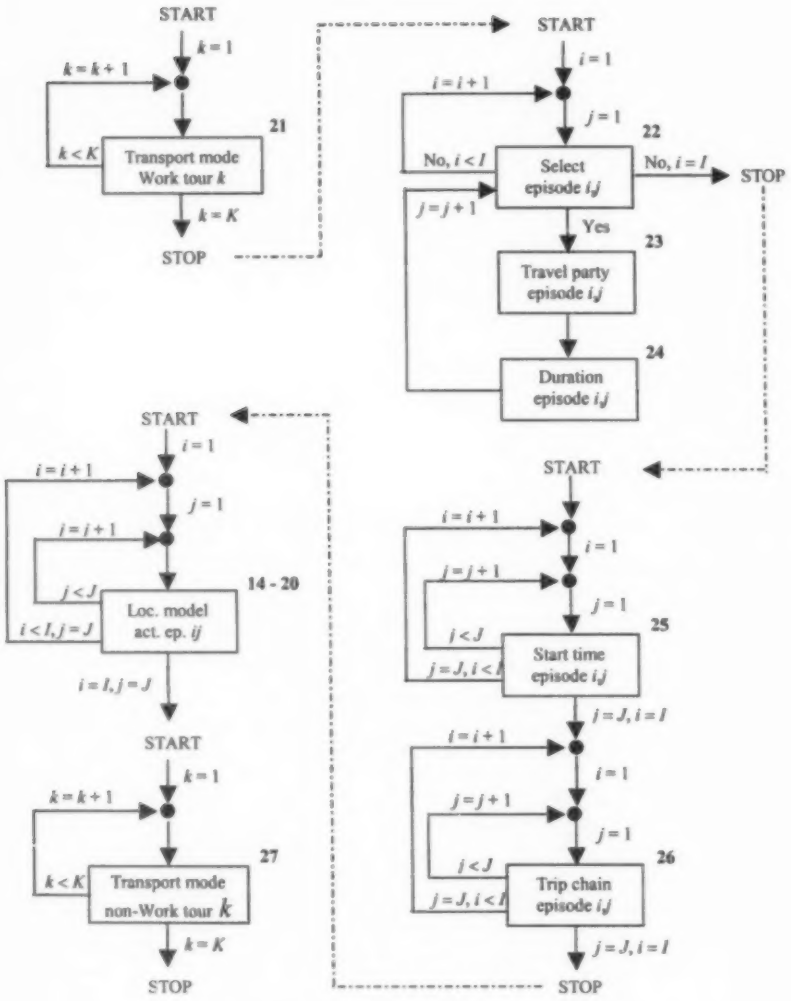

FIG. 3. The Process Model for the Flexible Part of Schedules

mode. In this way, we may expect that distributions of key schedule characteristics will be accurately predicted. Second, the sequencing of identified key choices is based on assumptions regarding the priority of decisions. For that reason, the skeleton is determined before flexible parts of the schedule and start time decisions before location and mode decisions, etc.

As an important distinct feature, the skeleton model component determines activity patterns on a continuous time scale (Figures 1 and 2). It consists of several subprocesses including:

1. Determining the pattern of sleep activities (1, 2)
2. Determining the pattern of the primary work/school activity (3–8).
3. Determining the pattern of secondary, fixed activities (9–13).
4. Determining the location of each fixed activity episode (14–20).

The component concerned with flexible activities (Figure 3) consists of the following subprocesses:

1. Choice of transport mode for the primary work/school tour(s) (21)
2. Determining selection, travel party and duration of flexible activities (4–24)
3. Determining start time and trip chaining of activities (25–26)
4. Determining the transport mode of each tour in the schedule (27)
5. Determining the location of each flexible activity (14–20)

The same location model is used for flexible and fixed activities. Possible differences in decision rules can be captured by decision trees, as activity type is incorporated as one of the potential condition variables. In the remainder of this section we describe the location model in more detail given its relevance for the illustration.

The location component chooses locations in descending order of priority of fixed activities. For each activity, the choice set consists of all postcode areas[1] (PCAs) in the Netherlands. First, the model chooses the municipality and next a PCA within the chosen municipality. At both levels, the model determines a choice by increasingly narrowing down the choice set in a number of steps. For the choice of municipality at the highest level, the first decision tree determines whether the activity (episode) is conducted within or outside the home municipality of the individual. If the last option is chosen, the choice of a municipality follows from a choice of an *order* and distance band. Five orders are distinguished based on population size. First-order municipalities are municipalities with an above regional function, second-order municipalities have a regional function, third-order municipalities have a local core function, and fourth and fifth-order municipalities correspond to smaller towns. Given the order, the choice of a distance band follows. The combination of order and distance band tends to reduce choice sets strongly. If there are still multiple alternatives left, the model selects a municipality semi-randomly using a parameterized probability function in which less distant locations within the given distance band have a higher probability of being selected.

For the choice of a zone (i.e., a PCA) within the chosen municipality a similar logic is used. Again, locations—in this case PCAs—are classified into four order categories. This time, orders are defined in terms of relative size of employment within the chosen municipality. Taking the relative size implies that PCAs of the same order may differ in employment size between municipalities. Given the chosen criterion, first-order PCAs typically belong to the city center, second-order to a band around the

1. In the model, postcode areas are defined at the level of the first 4 digits. To give an indication of spatial resolution: The Netherlands counts a total of over 4000 4PCAs. Thus, given a population of roughly 16 million people, the average number of individuals per 4PCA is approximately 4000 people and the average size is approximately 880 ha.

center, and last-order locations to the periphery of the municipality (cores around a city or the edge of cities). The choice of a PCA then is the result of a choice of order and next distance band in the same way as before. Again, if there are multiple choice alternatives left a semi-random selection is made using a similar probability function as in the case of municipalities with parameters specifically estimated for PCA choice.

In sum, the model takes into account that location choice has a random component and, at the same time, correlates with order and distance. By using this approach, the observed distribution of locations across combinations of size and distance category should be accurately predicted. The structure of each decision tree separately should take care of reproducing correlation structures between independent variables within this distribution.

4.2. Data

The empirical derivation of the model is based on four pooled activity diary data sets that were collected on different moments in the period of 1997–2000 in different regions of the Netherlands. Although the data sets were collected for different research projects, the survey methods and diary instruments used were largely the same. In each survey, all household members older than 12 years were invited to fill out a diary for two consecutive days. Diaries were balanced across days of the week. Only the diaries of adult members (i.e., household heads) of households were used to derive the decision trees. As a result, the training set used consists of a total of 9985 person-days. The data set covers a wide range of variation within the Netherlands in terms of urban character and region, although areas of lower urban density and areas in the periphery of the country are less well represented.

A complementary data set described the nation-wide study area. It included road network data, sectorial employment data, population data, and parking data. Network data were used to calculate mode specific fastest route distances and travel times between PCAs. Public-transport system data were only available at a higher spatial aggregation level. In sum, the physical data allowed us to measure conditions in terms of travel times by time-of-day and mode, travel costs per mode, parking space and costs, and availability and size of facilities by activity type. The impact tables described in the section that follows are based on the training data.

4.3. Illustration of impact measures

To illustrate the impact measures, we consider here location and mode-choice decision trees as examples. The location model predicts for each out-of-home activity the location where the activity is to be conducted. As implied by the process model, this involves seven decision trees (see Figure 2). Arbitrarily, we focus here on the choice between home municipality and other and, for cases where 'other' is chosen, the decision whether or not the nearest municipality of the chosen order is chosen. These two decisions largely determine the choice of making a short, local trip or a longer, above-local trip. Transport mode decisions involve two decision trees, one for the choice of a mode for each work/school trip and one for the choice of a mode for each tour for non-work/school activities.

4.3.1. Condition variables and model performance. Tables 1–5 show the condition variables used in deriving the decision trees. The choice of condition variables reflects the availability of physical data, the stage of development of the schedule in the concerned step in the decision process, and theories of which variables may moderate decisions for the concerned facet. The first table represents household and individual attributes that were used for all decision trees. The second table represents the condition variables that were used for the location decision trees considered here. The last three tables show the general set and specific sets used for the transport-mode decision trees. In addition to socio-economic variables, the tables generally include

TABLE 1
Socio-economic and Situational Attributes Used in all Decision Trees

Label	Definition	Levels
Urb	Urban density	0: highest density, 4: lowest density
Day	Day of the week	0: Monday, 6: Sunday
Comp	H composition	0: single 0 worker, 1: single 1 worker, 2: double 1 worker, 3: double 2 workers, 4: double 0 workers
Child	Children category	0: no children, 1: younger 6, 2: 6–12, 3: older 12
Age	Age category	0: < 25, 1: 25–44, 3:45–64, 4: > 64
SEC	Socio-economic class	0: low (below modal), 3: high (two times modal or more)
Ncar	# of cars in household	0: no cars, 1: 1 car, 2: 2 or more cars
Driver	*I* has driving license	0: is not a driver, 1: is driver
Gend	Gender of *I*	0: female, 1: male
Wstat	Work status of *I*	0: no work, 1: < 32 hours week, 2: 32 hours or more week
Pwstat	Work status of *P*	0: no work, 1: < 32 hours week, 2: 32 hours or more week

NOTE: *H* := concerned household; *I* := concerned individual; *P* := partner of concerned individual

TABLE 2
Condition Variables for Choice of Municipality (DT 15, 17)

Label	Definition	Categories
Adur	Duration (min.) of activity under concern	0: <0, 5], 1: <5,15], 2: <15,60], 3:<60,209], 4: <209,459], 5: > 459
Act	Type of activity under concern	0: work/school (long), 1: work/school (short), 2: bring/get, 3: vol. work, 4: other fix, 5: day shop., 6: service, 7: n.-day shop., 8: social, 9: leisure
Wdu	Total duration of work/school activity	0: <0, 385], 1: <385,490], 2: <490,525], 3:<525,560], 4: > 560
Orde0	Order of home municipality	0: lowest … 4: highest
Orde1[2]	Order of chosen municipality	0: lowest … 4: highest
Orderb*i*	Maximum order across municipalities available in distance band *i*	0: lowest … 4: highest
Nearb*i*[1]	Nearest band for municipality of order 5-*i*	0: first band … 4: fifth band
Avb*i*[2]	Chosen order is available in band *i*	0: no, 1: yes
AvailT	Width (min.) of time window for activity	0: <0, 130], 1: <130,270], 2: <270,415], 3:<415,654], 4: > 654
Aflex	Activity under concern is flexible type	0: no, 1: yes
Avgem	Another municipality is available in time window for activity	0: no, 1: yes
TRpucai[2]	Travel time ratio (%) public-transport/car to nearest mun. of order *i* = 3..5	0: <0, 171], 1: <171,201], 2: <201,232], 3:<232,249], 4: > 249
CRpucai[2]	Travel cost ratio (%) public-transport/car to nearest mun. of order *i* = 3..5	0: <0, 110], 1: <110,150], 2: <150,175], 3:<175,228], 4:<228,400], 5: > 400
Cpark*i*[2]	Average hour tariff of long parking of nearest mun. of order *i* = 3..5	0: <0,0], 1: <0,19], 2: <19,32], 3:<32,92], 4:<92,275], 5: > 275
Avail*i*[2]	Municipality of chosen order is available in band *i* given time window	0: no, 1: yes

NOTES: [1]Included only in DT 15; [2]Included only in DT 17

TABLE 3
Condition Variables Generally Used for Transport Mode Choice (DT 21, 27)

Label	Definition	Categories
TRpubi	Travel time ratio Public, Bike to loc.	0: <0, 1.0], 1: <1.0,1.2], 2: <1.2,1.5], 3: <1.5,1.84], 4:<1.84,2.48], 5: > 2.48
TRcoff	Car travel time ratio in-peak, out-peak to loc.	0: <0, 1.0], 1: <1.0,1.53], 2: <1.53,2.15], 3: <2.15,2.87], 4:<2.87,3.87], 5: > 3.87
TRpuca	Travel time ratio Public, Car to loc.	0: <0, 0.73], 1: <0.73,1.0], 2: <1.0,1.56], 3: <1.56,2.64], 4:<2.64,4.21], 5: > 4.21
CRpuca	Cost ratio Public, Car to loc.	0: <0, 1.1], 1: <1.1,1.5], 2: <1.5,1.75], 3: <1.75,2.28], 4:<2.28,4.0], 5: > 4.0
TRvona	Ratio access + egress time, total Public travel time (%)	0: <0, 18], 1: <18,25], 2: <25,32], 3: <32,40], 4:<40,46], 5: > 46
PRbeta	Ratio of paid parking places (%)	0: <0, 9], 1: <9,17], 2: <17,23], 3: <23,62], 4: > 62
Cpark	Average parking price per hour (fl cent)	0: <0, 0], 1: <0,19], 2: <19,32], 3: <32,92], 4: > 92

TABLE 4
Condition Variables for Transport Mode Choice for Work/School Trips (DT 21)

Label	Definition	Categories
Wdu	Total duration of work/school activity in *S* (min.)	0: <0, 385], 1: <385,490], 2: <490,525], 3:<525,560], 4: > 560
Wdutot	Total duration of work/school and voluntary work activity in *S* (min.)	0: <0, 240], 1: <240,360], 2: <360,480], 3: > 480
Nsec	# Of out-of-home non-work activities in *S*	0: 0, 1: 1, 2: 2, 3: [3,4], 4: 5, 5: > 5
Bget	Bring/get activity included in *S*	0: no, 1: yes
Pwdu	Total duration of work/school activity in *S* of *P* (min.)	0: <0, 385], 1: <385,490], 2: <490,525], 3:<525,560], 4: > 560
Pwdutot	Total duration of work/school and vol. work activity in *S* of *P* (min.)	0: <0, 240], 1: <240,360], 2: <360,480], 3: > 480
Pnsec	# Of out-of-home non-work activities in *S* of *P*	0: 0, 1: 1, 2: 2, 3: [3,4], 4: 5, 5: > 5
Pbget	Bring/get activity included in *S* of *P*	0: no, 1: yes
inAM	Travel to work/school in morning peak	0: no, 1: yes
inPM	Travel back home in evening peak	0: no, 1: yes
WduT	Duration of work/school activity in *T* (min.)	0: <0, 300], 1: <300,443], 2: <443,495], 3:<495,525], 4:<525,555], 5: > 555
Cadist	Car distance to location (km)	0: <0, 3.1], 1: <3.1,6.4], 2: <6.4,12.7], 3:<12.7,22.0], 4:<22.0,45.8], 5: > 45.8
Trcon	There is a train connection to loc.	0: no, 1: yes
Nwloc	# Of different work locations on *T*	0: 1, 1: > 1
Avo	Activity before work/school	0: none, 1: bring/get, 3: other
Ana	Activity after work/school	0: none, 1: bring/get, 3: other
Pwo	*P* has work overlapping in time with *T*	0: no, 1: yes
Pbget	*P* has bring/get overlapping in time with *T*	0: no, 1: yes
Pnact	# Of fixed activities of *P* during *T*	0: 0, 1: 1, 2: > 1
Pdist	Longest car distance across activities of *P* overlapping in time with *T* (km)	0: <0, 0], 1: <0,2.71], 2: <2.71,5.76], 3: <5.76,14.18], 4:<14.18,34.95], 5: > 34.95
Inuse	Car in use by P for work trip by previous decision	0: no, 1: yes

NOTES: *S* is current schedule, *P* is partner of the individual; *T* is the work-based tour under concern

TABLE 5
Condition Variables for Transport Mode Choice for Non-work Tours (DT 27)

Label	Definition	Categories
Avcar	Availability of a car in time window	0: no, 1: yes, 2: unknown
Nsec	# Of out-of-home non-work activities in *S*	0: 0, 1: 1, 2: 2, 3: [3,4], 4: 5, 5: > 5
Wdu	Total duration of work/school activity in *S* (min.)	0: <0, 385], 1: <385,490], 2: <490,525], 3:<525,560], 4: > 560
Wdutot	Total duration of work/school and voluntary work activity in *S* (min.)	0: <0, 240], 1: <240,360], 2: <360,480], 3: > 480
BTchain	Episode of the day start time of *T*	0: 1-st episode .... 5: last episode
Aty1	Type of first activity in *T*	0: work/school (short), 1: bring/get, 2: vol. work, 3: other fix, 4: day shop., 5: service, 6: n.-day shop., 7: social, 8: leisure
Aty2	Type of second activity in *T*	0: work/school (short), 1: bring/get, 2: other fix, 3: {day shop., service, n.-day} shop., 4: {social, leisure}
Adur1	Duration class of first activity in *T*	0: short, 1: medium, 2: long
Cbrget	There is a bring/get activity in *T*	0: no, 1: yes
Cgroc	There is a daily shop activity in *T*	0: no, 1: yes
Cserv	There is a service activity in *T*	0: no, 1: yes
Cshop	There is a non-daily shop activity in *T*	0: no, 1: yes
Csoco	There is a social activity in *T*	0: no, 1: yes
Cleiso	There is a leisure activity in *T*	0: no, 1: yes
Cnlout	There is a work or vol. work act. in *T*	0: no, 1: yes
Cadist	Car distance of longest link in *T* (km)	0: <0, 0], 1: <0,2], 2: <2,4], 3: <4,6], 4:<6,10], 5:<10,30], 6:<30,80], 7: > 80
Pbrget	*P* has bring/get activity overlapping in time with *T*	0: no, 1: yes
Pserv	*P* has service activity overlapping in time with *T*	0: no, 1: yes
Pstat	Status of *T* relative to partner	0: unknown, 1: no overlap, 2: overlap no car, 3: overlap car trip
Pdist	Distance (km) of tour of *P* overlapping with *T*	0: unknown, 1: no such tour, 2: 0, 3: [1, 10>, 4: [10, 45>, 5: ≥ 45

accessibility indicators, travel costs variables and potentially relevant characteristics of the current schedules of the individual and partner, if any. In particular, the relative attractiveness of the transport modes is indicated by travel-time and travel-costs ratios between modes and availability of parking facilities (Table 3). All continuous variables were discretisized using the equal-frequency interval method.

A CHAID-based decision tree induction method was used to derive the decision trees, whereby the entire sample was used for training. Table 6 shows some statistics of the decision trees that are considered here. Successively the columns represent the minimum number of cases per leaf node (nmin), which is a parameter of the induction method, the number of condition variables (ncond), the number of choice alternatives (nobs), the number of leaf nodes of the resulting tree (nleafs) and several measures of goodness-of-fit. The first two fit measures express an expected ratio of correct predictions under a probabilistic action-assignment rule of type (5) for a null model and the derived model respectively. Hereby, the null model is defined as the tree consisting of the root node only. So, this model only uses the overall sample choice frequency distribution for prediction. e_{incr} Is a combination of the first two measures and indicates the increase in fit on a zero-one scale.

As Table 6 indicates, the decision trees all perform satisfactory in terms of prediction accuracy (e) and yield a substantial improvement in fit compared to the null

model (e_{incr}). Furthermore, the number of leaf nodes gives an indication of the size of the trees. The largest tree has 105 leaf nodes and the smallest tree still has 51 leaf nodes. In general, tree size tends to be correlated with the number of cases available for training, on the one hand, and the minimum group size setting (nmin), on the other. Sizes as we find here are not exceptional and trees would even have been bigger if survey data of a national-scale had been used. This clearly demonstrates the problem of interpreting decision trees in terms of impacts of condition variables on choice.

4.3.2. Location choice. Tables 7 and 8 display the impact tables for the location decision trees considered here. In these tables as well as in the impact tables considered in the next section, we used a Chi-square measure of fit as a specification of D in equations (9) and (10). The rank order of variables on this measure can be interpreted as a rank order of impact on the choice variable overall (the first *IS* column) or per level of the choice variable (the next *IS* columns). Moreover, the size of the difference be-

TABLE 6
Some Statistics of the Decision Trees Used as Examples

DT id	DT label	nmin	ncond	nalt	nobs	nleaf	e_0	e	e_{incr}
15	L mun. in/out	90	29	2	18758	105	0.512	0.625	0.277
17	L mun. nearest	79	38	2	7932	55	0.503	0.727	0.451
21	Mode to work	36	39	4	3665	51	0.381	0.590	0.338
27	Mode to non-work	90	38	4	9523	56	0.425	0.607	0.317

TABLE 7
Impact Table: Choice Between Home Municipality and Other (DT 15)

Condition	*IS*	*IS*(home)	*IS*(other)	*MS*(home)	*MS*(other)
Urb	11.69	4.94	6.76	−0.54	0.54
Comp	14.15	5.97	8.18	−0.17	0.17
Child	33.48	14.22	19.26	0.04	−0.04
Day	12.87	5.45	7.42	−0.64	0.64
Age	0.14	0.06	0.08	0.04	−0.04
SEC	10.45	4.39	6.06	−0.44	0.44
Near	4.71	2.00	2.71	−0.56	0.56
Gend	62.00	26.26	35.73	−1.00	1.00
Driver	2.46	1.04	1.43	−1.00	1.00
Wstat	27.93	11.74	16.20	−0.88	0.88
Pwstat	0.16	0.07	0.10	1.00	−1.00
Adur	121.26	51.12	70.13	−1.00	1.00
Act	18911.43	7676.01	11235.31	0.18	−0.18
Wdu	5.00	2.15	2.86	−1.00	1.00
Orde0	4857.42	1864.81	2992.63	0.92	−0.92
Orde1	0	0	0		
Orderb1	0	0	0		
Orderb2	0.11	0.05	0.07	−1.00	1.00
Orderb3	1.04	0.44	0.60	−1.00	1.00
Orderb4	0	0	0		
Orderb5	0	0	0		
Nearb1	0	0	0		
Nearb2	46.65	19.80	26.83	−1.00	1.00
Nearb3	1.48	0.63	0.86	1.00	−1.00
Nearb4	0	0	0		
Nearb5	0	0	0		
AvailT	103.63	43.56	60.08	0.70	−0.70
Aflex	1.68	0.71	0.97	1.00	−1.00
Avgem	12.97	5.58	7.39	1.00	−1.00

TABLE 8
Impact Table: Choice Between Nearest Municipality and Other (DT 17)

Condition	*IS*	*IS*(near)	*IS*(other)	*MS*(near)	*MS*(other)
Urb	40.57	22.21	18.37	0	0
Comp	0	0	0		
Child	5.29	2.83	2.46	1.00	−1.00
Day	8.20	4.41	3.79	−1.00	1.00
Age	0.29	0.16	0.13	0.33	−0.33
SEC	32.89	17.42	15.46	−1.00	1.00
Near	1.59	0.85	0.73	1.00	−1.00
Gend	4.32	2.33	1.99	−1.00	1.00
Driver	0	0	0		
Wstat	0.58	0.31	0.27	−1.00	1.00
Pwstat	0.22	0.12	0.10	−1.00	1.00
Adur	19.93	10.75	9.18	−1.00	1.00
Act	0	0	0		
Wdu	0.38	0.21	0.17	−1.00	1.00
Orde0	99.93	56.85	43.08	−0.52	0.52
Orde1	7224.67	4737.97	2486.66	0.74	−0.74
Orderb1	653.28	412.79	240.48	0.15	−0.15
Orderb2	9.98	5.39	4.58	1.00	−1.00
Orderb3	1.74	0.94	0.80	0.04	−0.04
Orderb4	2.21	1.21	1.01	1.00	−1.00
Orderb5	0	0	0		
Avb1	94.68	53.91	40.77	1.00	−1.00
Avb2	0	0	0		
Avb3	0.77	0.42	0.36	−1.00	1.00
Avb4	0	0	0		
Avb5	0	0	0		
AvailT	1.01	0.55	0.47	−1.00	1.00
Aflex	1.44	0.77	0.66	1.00	−1.00
avgem	0	0	0		
TRpuca3	0	0	0		
TRpuca4	15.60	8.49	7.11	−1.00	1.00
TRpuca5	4.89	2.66	2.24	0.01	−0.01
CRpuca3	25.48	13.77	11.70	0.08	−0.08
CRpuca4	580.03	305.22	274.80	0.78	−0.78
CRpuca5	0	0	0		
Cpark3	0.00	0.00	0.00	1.00	−1.00
Cpark4	0	0	0		
Cpark5	0	0	0		

tween condition variables on this measure can be interpreted as the size of the difference in impact. One should keep in mind however that the difference in Chi-square represents a *squared* difference in relative impacts. Finally, the *MS* columns give the degree of monotonicity of the impacts per level of the choice variable.

First, we consider the choice between conducting the activity under concern within (*home*) or outside (*other*) the own municipality of the individual (Table 7). In this case, the choice variable is a binary variable, so that the measures are strongly correlated between choice alternatives on the *IS* as well as the *MS* dimension. The reason for this is that in each case the probability of the first alternative equals one minus the probability of the second alternative. For that reason the correlation between MS_1 and MS_2 is even perfect: for each condition variable the measures have the same absolute value and opposite signs. Also in the case of *IS*, one may expect a perfect correlation since an increase in frequency of the first alternative equals the decrease in frequency of the second alternative. However, the Chi-square measure used here does not consider absolute differences but differences relative to an expected frequency of the choice alternative under concern. It is for that reason that measured impacts for *Home*, of which the base frequency is 57.7 %, is consistently lower than for *Other*, of which the base frequency is 42.3 %. Because we consider

this property of Chi-square as appropriate, IS_1 or IS_2 is redundant in the impact table only in the sense that either one of the two can be found by the relationship $IS = IS_1 + IS_2$.

Now looking at the contents of the table, we see that activity type (*Act*, a ten-level variable) is by far the most important variable for this choice. The monotonocity index *MS* in this case is close to zero (±0.18) as could be expected given the fact that activity type is a nominal variable and there is no information in the order of the levels of the variable. The second most important variable is order of the home municipality (*Orde0*). This variable is measured on an ordinal scale whereby a higher value indicates a higher order. The high monotonicity index of 0.92 for *Home* indicates that choice of *Home* almost monotonically increases with order, which is what one could expect. The third most influential variable, i.e. the width of the available time window for the activity (*availT*), has an impact of a lower order of magnitude. Here the monotonocity index for *Home* is also positive implying that the choice of *Home* increases with increasing time availability. An explanation of this counter-intuitive relationship can be found by examining the underlying decision tree. As it appears, the direction of the impact varies depending on the specific branch of the tree (i.e., the context) in which the condition variable plays a role. A negative impact seems to be related to opportunities of linking the activity to another higher-priority activity, especially work, that takes place outside the home municipality and that, at the same time, tends to reduce the available time window for the activity. So, the negative relationship may be explained by a wish to combine the activity with another activity at a distant location. Activity duration is the fourth most important variable with a perfectly monotonously decreasing effect on *Home* choice, as one would expect.

As for the socio-demographic variables, a number of observations are relevant. First, all variables have some impact. In order of decreasing importance we find gender, children, work status, day of the week, household composition (*Comp*), income (*SEC*), driver, work status (*Wstat*), number of cars (*Ncar*) and driving license (*Driver*). For binary variables (Gender and Driver) and ordinal variables (*Day*, *SEC*, *Ncar*) the monotonocity index has a meaning and the outcomes are well interpretable. For example, being male decreases the probability of choosing *Home*, as do higher income and later days in the week.

Finally, condition variables of the spatial structure of activity locations also have an impact. We already mentioned the order of the home municipality, which is by far the most important one of this category. Urban density of the home location has a negative impact. This is what one would expect given the fact that the variable is coded such that a lower value indicates higher density. The accessibility variables (*Orderb*1-5 and *Nearb*1-5) tend to have very small impacts. Only the nearest band for an order 2 municipality has a meaningful impact on *Home* choice in the anticipated direction.

The second example (Table 8) concerns the choice of a municipality in cases where *Other* was chosen in the previous step. The decision occurs after the choice of a municipality order and determines whether or not the nearest municipality of that order is chosen. A first observation is that overall the impacts are slightly lower than in the previous case. The overall size of impacts is determined by the combination of two properties of the decision tree, namely the size of the sample and the increase in goodness-of-fit achieved by the tree (e_{incr}). Chi-square values increase with both sample size and the fit measure. In this case, the forces work in opposite direction, as the decision tree has a lower number of training cases (7932 vs. 18758) and a higher fit (0.560 vs. 0.468). As it appears, the effect of sample size is bigger than the effect of fit, in this case.

Turning to the contents of Table 8, the chosen order for the municipality (*Orde1*) is by far the most important variable for the decision. The choice of nearest increases almost monotonically (*MS* = 0.74) with increasing order, which is what one would ex-

pect. The second most important variable is the order of the nearest band (*Orderb1*). There is a tendency ($MS = 0.15$) of increasing probability of *Near* with increasing levels on this variable. Interestingly, the next most influential variable is the cost ratio between public transport and car for reaching the nearest municipality of the second highest order. The higher this ratio the higher the predicted probability of *Near* ($MS = 0.78$). This suggests that the relative costs of public transport acts as a proxy for accessibility of high order municipalities. In addition, the availability of a municipality of the chosen order in the nearest distance band (*avb1*) has an impact with anticipated sign ($MS = 1$).

With respect to the relative impacts of different types of condition variables, several observations are relevant. First, socio-economic variables have relatively small impacts in this stage of the decision process. Income level (*SEC*) stands out as the socio-economic variable with the biggest impact. The choice of *Near* decreases monotonically ($MS = -1$) with increasing income level. The same holds for having children (*Child*) and being male, be it that the impacts of these variables are smaller. Second, the relative attractiveness of transport mode measured by costs ratios and travel time ratios have impact. We already mentioned the cost ratio between public transport and car, which has the biggest impact. In addition, the relative travel time by public transport to the nearest fourth-order municipality has a monotonically decreasing impact on *Near*. Finally it is worth noting that parking prices do not affect predictions at this level, except that the choice of nearest municipality decreases with increasing parking tariff in nearest order of 3.

4.3.3. Mode choice. In Albatross, mode choice is made separately for primary work/school trips and tours for other activities. In both cases, the mode alternatives considered include slow (non-motorized), car driver, public transport, and car passenger. Impact tables related to these decision trees are shown in Tables 9 and 10 and can be interpreted as follows.

First, as it appears from Table 9, distance to the work/school location is by far the most influential variable with expecting monotonicity levels for the different mode alternatives. Car passenger (*Pass*) undergoes a less monotonous impact across levels of this variable. The next most important variables are number of cars in the household (*Ncar*), ratio paid/free parking places (*PRbeta*) and having a driving license (*Driver*) with high levels and expected directions of monotonicity. All these variables monotonically increase the choice of car driver and decrease the choice of public transport in the model.

When we look at the differential impacts of types of condition variable, we see that socio-economic variables have only a limited impact here, whereas attributes of the transportation system have a relative big impact. Especially, travel time ratios and parking tariffs for the chosen destination emerge with substantial impacts. Interestingly, the capacity ratio of paid/free parking places (*Pbeta*) is the most important variable in this category and monotonically decreases the choice of car and increases the choice of public transport. Height of parking tariff (*Cpark*) has a similar but smaller impact. Furthermore, the access/egress time ratio in public transport services (*TRvona*) tends to decrease the choice of public transport, but the impact is small. Finally, it is worth noting that congestion on routes to the work location is influential. The travel time ratio between congested and free floating conditions (*TRcoff*) both decrease the choice of car and increase the choice of public transport monotonically.

Second, also for tours for other purposes than work/school (Table 10) distance (*Dist*) is the primary moderator of mode choice followed (at a distance) by number of cars in the household (*Ncar*). As for socio-economic variables, Gender (*Gend*) emerges as the strongest moderating factor in the model. Being male increases the choice of car and decreases the choice of a slow mode. The opposite impacts of work status of the individual and that of his/her partner suggest an impact of traditional

TABLE 9

Impact Table: Choice of Transport Mode Work/School Trips (DT 21)

Condition	*IS*	*IS*(slow)	*IS*(car)	*IS*(pub)	*IS*(pass)	*MS*(slow)	*MS*(car)	*MS*(pub)	*MS*(pass)
Urb	10.19	0.24	3.09	6.72	0.13	−1.00	1.00	−1.00	0.65
Comp	9.53	0.02	1.45	7.48	0.58	0.00	0.00	0.00	0.00
Child	52.07	30.30	19.27	0.24	2.26	0.18	−0.11	−1.00	−0.22
Day	0.00	0.00	0.00	0.00	0.00				
Age	0.00	0.00	0.00	0.00	0.00				
SEC	0.00	0.00	0.00	0.00	0.00				
Ncar	968.40	521.03	398.11	45.25	4.00	−1.00	1.00	−1.00	−1.00
Gend	1.01	0.45	0.00	0.51	0.04	1.00	−1.00	−1.00	−1.00
Driver	716.44	13.97	335.47	366.94	0.07	−1.00	1.00	−1.00	1.00
Wstat	0.00	0.00	0.00	0.00	0.00				
Pwstat	2.95	0.23	0.60	0.01	2.11	0.00	0.00	0.00	0.00
Wdu	0.80	0.26	0.36	0.05	0.12	−1.00	1.00	−1.00	−1.00
Wdutot	0.00	0.00	0.00	0.00	0.00				
Nsec	0.00	0.00	0.00	0.00	0.00				
Bget	3.91	0.29	1.31	0.64	1.67	−1.00	1.00	−1.00	−1.00
Pwdu	0.00	0.00	0.00	0.00	0.00				
Pwdutot	0.00	0.00	0.00	0.00	0.00				
Pnsec	0.00	0.00	0.00	0.00	0.00				
Pbget	2.93	0.00	0.44	0.27	2.22		−1.00	1.00	1.00
inAM	0.00	0.00	0.00	0.00	0.00				
inPM	0.00	0.00	0.00	0.00	0.00				
Wdu	0.00	0.00	0.00	0.00	0.00				
Cadist	7333.68	4709.25	804.59	1684.73	144.13	−1.00	0.82	1.00	−0.18
Trcon	3.33	0.03	0.40	0.02	2.88	−1.00	1.00	−1.00	−1.00
TRpubi	7.94	4.40	2.94	0.08	0.52	−0.50	0.57	−0.40	−0.60
TRcoff	31.38	0.17	6.71	23.42	1.07	0.26	−1.00	1.00	−0.59
TRpuca	0.00	0.00	0.00	0.00	0.00				
CRpuca	4.58	0.29	1.70	2.59	0.00	−1.00	1.00	−1.00	−1.00
TRvona	2.69	0.92	0.66	0.16	0.95	0.50	0.05	−1.00	−1.00
PRbeta	727.24	7.55	253.48	458.37	7.84	0.31	−1.00	1.00	−0.10
Cpark	108.80	1.61	24.67	82.19	0.33	−0.08	−1.00	1.00	−0.75
Nwloc	0.00	0.00	0.00	0.00	0.00				
Avo	0.00	0.00	0.00	0.00	0.00				
Ana	0.00	0.00	0.00	0.00	0.00				
Pwo	2.78	1.60	1.12	0.01	0.05	1.00	−1.00	1.00	1.00
Pbget	0.00	0.00	0.00	0.00	0.00				
Pnact	2.85	0.56	0.87	0.02	1.40	1.00	−1.00	1.00	1.00
Pdist	0.00	0.00	0.00	0.00	0.00				
Inuse	0.00	0.00	0.00	0.00	0.00				

TABLE 10
Impact Table: Choice of Transport Mode Non-work/school Tours (DT 27)

Cond	Chisq	Chiwa	Ccar	Cpub	Ccarpl	Rhiwa	Rcar	Rpub	Rcarpl
Urb	7.13	1.04	0.76	1.51	3.83	−0.13	0.44	−0.86	−0.06
Comp	0.69	0.23	0.43	0.03	0.00	0.33	−0.33	0.33	0.33
Child	0.00	0.00	0.00	0.00	0.00				
Day	0.00	0.00	0.00	0.00	0.00				
Age	6.80	0.22	1.61	0.03	4.94	0.20	−1.00	1.00	1.00
SEC	0.26	0.09	0.17	0.00	0.01	0.00	0.00		0.00
Ncar	2158.20	845.92	1146.02	110.51	55.74	−1.00	1.00	−0.97	1.00
Gend	426.32	10.34	148.09	3.56	264.34	−1.00	1.00	−1.00	−1.00
Driver	7.16	3.00	3.82	0.08	0.26	−1.00	1.00	−1.00	1.00
wstat	11.35	3.91	6.95	0.01	0.48	−1.00	1.00	1.00	−0.78
Pwstat	3.33	1.46	1.81	0.01	0.05	1.00	−1.00	−1.00	−1.00
Avcar	0.00	0.00	0.00	0.00	0.00				
Nsec	1.75	0.59	1.12	0.02	0.03	−1.00	1.00	−1.00	−1.00
Wdu	0.00	0.00	0.00	0.00	0.00				
Wdutot	0.00	0.00	0.00	0.00	0.00				
Btchain	0.00	0.00	0.00	0.00	0.00				
Aty1	0.02	0.00	0.00	0.00	0.00				
Aty2	0.00	0.00	0.00	0.00	0.00				
Adur1	50.58	17.17	6.23	3.31	23.87	−1.00	0.66	0.82	1.00
Cbrget	0.68	0.24	0.09	0.00	0.35	1.00	−1.00		−1.00
Cgroc	0.00	0.00	0.00	0.00	0.00				
Cserv	0.00	0.00	0.00	0.00	0.00				
Cshop	0.00	0.00	0.00	0.00	0.00				
Csoco	0.00	0.00	0.00	0.00	0.00				
Cleiso	0.00	0.00	0.00	0.00	0.00				
Cnlout	0.00	0.00	0.00	0.00	0.00				
Cadist	25167.09	15507.72	3608.72	2858.38	3192.44	−1.00	0.82	1.00	0.88
CRpuca	0.00	0.00	0.00	0.00	0.00				
TRpubi	0.00	0.00	0.00	0.00	0.00				
TRcoff	0.00	0.00	0.00	0.00	0.00				
TRpuca	0.00	0.00	0.00	0.00	0.00				
TRvona	19.31	8.76	6.69	0.18	3.68	1.00	−1.00	−1.00	−1.00
PRbeta	285.48	23.30	46.73	182.44	33.01	1.00	−1.00	1.00	−1.00
Cpark	136.97	51.28	60.26	7.13	18.29	1.00	−1.00	1.00	−1.00
Pbrget	0.00	0.00	0.00	0.00	0.00				
Pserv	0.00	0.00	0.00	0.00	0.00				
Pstat	57.40	1.65	16.35	0.64	38.75	0.60	−0.16	−0.29	0.01
Pdist	12.04	0.08	0.67	1.99	9.29	−1.00	−1.00	−1.00	1.00

role patterns within households. Of the cost and travel time variables the ratio paid/free parking (*Pbeta*) is again the most important variable with perfect monotonous impacts in expected directions. Height of parking tariffs (*Cpark*) is the second most important with impacts in the same directions. Access/egress time ratio of total public transport time (*TRvona*) works in favor of slow mode and reduces public transport choice. Finally, it is worth noting that the activity pattern of the partner (if any) has some impact here. In cases where the partner makes a tour overlapping in time (*Pstat*) the choice of car decreases and the choice of a slow mode increases.

5. CONCLUSIONS AND DISCUSSION

Tree induction methods have been developed to explore the complexity in data of observations. In the context of space-time data, these methods have been applied to extract decision rules from activity-travel diaries either as an explorative tool or as a means of calibrating a rule-based computational process model. Although these models claim to provide a more detailed insight into context-dependent spatial choice behavior, these rule-based models may be criticized as being too much of a black box. Unlike algebraic models such as the gravity model and the competing destinations models, rule-based models do not give elasticities, and hence are difficult to interpret.

The present paper presented a possible solution to this problem. The concept of impact tables has been introduced and a simple method of deriving such tables has been developed and illustrated in the context of the Albatross model system. The measures that were introduced indicate for each condition variable and each decision tree or rule-set in general the sensitivity of the predicted choice for levels of the condition variable. In addition, an index of monotonicity was developed to obtain information on the direction of the impacts. As illustrated by location and mode choice decision trees derived for the Albatross model, the impact table provides substantial information for theory development and policy making. The impact table allows one to rank condition variables and assess the difference in impact between condition variables, so that the most relevant variables for policy making can be identified. For example, the results of the case study suggests that parking price policies and transport-system service level policies are effective moderators of transport mode choice, whereas the spatial structure of locations for activities is more influential for the choice of activity locations. Such observations are particularly meaningful in the context of an activity-based model, which takes into account a large set of activity schedule and household context variables that may constrain and moderate location or mode choices.

Besides advantages we should point to a potential limitation of the approach. It has been found that the structure of empirically derived decision trees tends to be sensitive to the selection and way of operationalizing of variables in applications where the variables are strongly intercorrelated (Quinlan 1993). Depending on such choices of the analyst, certain condition variables may or may not be selected by the induction method in developing the tree. As a consequence, one could obtain varying values of impacts of condition variables. This observation does not degrade the method developed here, as it explicitly intends to measure the sensitiveness of a given model and not the generalizability of its predictions. However, making elasticities explicit makes one more aware of the problem. This problem is not unique for rule-based models, but is closely related to the problem of multi-collinearity in estimating parametric choice models. Furthermore, instability of decision tree structures tends to be limited to those condition variables that make only a small unique contribution to prediction accuracy. This means that big elasticities tend to be more stable and hence better generalizable across model specifications. Using an appropriate feature selection method could increase stability of rule-based models.

LITERATURE CITED

Arentze, T., A. Borgers, F. Hofman, S. Fujii, C. Joh, A. Kikuchi, R. Kitamura, H. Timmermans, and P. Van der Waerden (2001) Rule-based versus Utility-maximizing Models of Activity-Travel Patterns: A Comparison of Empirical Performance. In *Travel Behaviour Research: The Leading Edge*, by D. Hensher, pp. 569–584. Amsterdam: Pergamon.

Arentze, T., F. Hofman, H. van Mourik, and H. Timmermans (2002) The Spatial Transferability of the Albatross Model System: Empirical Evidence from Two Case Studies. The 81th Transportation Research Board Meeting, January 11–16, Washington D.C. *Transportation Research Record* (forthcoming)

Arentze, T.A., and H.J.P. Timmermans (2000) Albatross: *A Learning Based Transportation Oriented Simulation System.* In European Institute of Retailing and Services Studies, Eindhoven, pp. 152–168.

Arentze, T.A. and H.J.P. Timmermans (2002) Albatross 2.0. Eindhoven: European Institute of Retailing and Services Studies.

Arentze, T.A., Hofman, F., van Mourik, H., Timmermans, H.J.P., and Wets, G. (2000) Using Decision Tree Induction Systems for Modeling Space-Time Behavior. *Geographical Analysis* 32, 52–72.

Bowman, J.L., and M.E. Ben-Akiva (1999) The Day Activity Schedule Approach to Travel Demand Analysis. Paper presented at the 78th Annual Meeting of the Transportation Research Board, Washington, D.C., January 10–14.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone (1984). *Classification and Regression Trees*. Belmont (CA): Wadsworth.

Gahegan, M. (2000a) On the application of inductive machine learning tools to geographical analysis. *Geographical analysis* 32, 113–139.

Gahegan, M. (2000b) The case for inductive and visual techniques in the analysis of spatial data. *Journal of Geographical Systems* 2, 77–83.

Kass, G.V. (1980) An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics* 29, 119–127.

Kira, K., and L.A. Rendall (1992) A Practical Approach to Feature Selection. Machine Learning: Proceedings of the Ninth International Conference.

Kononenko, I. (1994) Estimating attributes: Analysis and extensions of relief. Proceedings of the European Conference on Machine Learning.

Moons, E., G. Wets, K. Verhoef, T. Arentze, and H. Timmermans (2002) The Impact of Irrelevant Attributes on the Performance of Classifier Systems in Generating Activity Schedules. Paper presented at the 81th Transportation Research Board Meeting, Washington D.C., January 13–17.

Piramuthu, S. (1996) Effects of feature selection with 'blurring' on neurofuzzy systems. In *Proceedings of the 7th International Workshop on Algorithmic Learning Theory*, pp. 135–142.

Quinlan, J. R. (1993) *C4.5 Programs for Machine Learning*. San Mateo (CA): Morgan Kaufmann Publishers.

Steinberg, D., and P. Colla (1997) *CART: Classification and regression trees*. San Diego (CA): Salford Systems.

Thill, J.C., and A. Wheeler (1999) Tree Induction of Spatial Choice Behavior. Paper presented at the RSAI meetings in Dublin, Ireland, August 1999.

Timmermans, H.J.P., T.A. Arentze, and C.H. Joh (2002) Analysing space-time behaviour: New approaches to old problems. *Progress in Human Geography* 26, 2, 175–190.

Wets, G., Vanhoof, K., Arentze, T.A., and Timmermans, H.J.P. (2000) Identifying Decision Structures Underlying Activity Patterns: An Exploration of Data Mining Algorithms. *Transportation Research Record* 1718, 1–9.

Yamamoto, T., R. Kitamura, and J. Fujii (2002) An Analysis of Driver's Route Choice Behavior by Data Mining Algorithms. Paper presented at the 81th Transportation Research Board Conference, Washington D.C., January 13–17.

Francesca Medda, Peter Nijkamp, and Piet Rietveld

Urban Land Use For Transport Systems and City Shapes

Urban land use is contingent upon the transportation system which is in turn dependent on business development in the city. The urban economics literature influenced by the Solow-Vickrey model has extensively studied the above relationship. In the present paper we re-examine the Solow-Vickrey model of the long narrow city by defining a new general mathematical formulation in which the urban shape is a function of the allocation of land for transport. We focus in particular on emergence conditions for different types of polycentric cities. The methodology discussed in the paper provides a new analytical framework for assessing the impact of urban transport flows on city growth.

INTRODUCTION

Structural and qualitative transformations in cities are often due to extensions of and new developments in transport infrastructure, in particular roads (Wang 2000; Martinez and Araya 2000; Peeters, Thisse, and Thomas 1998; Wheaton 1998; Cervero and Wu 1997; Berry and Kim 1993; Ladd and Wheaton 1991). The rise in road traffic and the subsequent increase in land use devoted to road infrastructure has produced many externalities that directly and indirectly affect the morphology of our cities.

The increase in urban land earmarked for transport systems is often initiated by the need to alleviate the congestion problem in urban areas. Travel time, and subsequently travel cost, increases when the movement from an origin to a destination is influenced by other travellers using the same infrastructure. More precisely, congestion occurs when the capacity of the transport infrastructure is insufficient to meet simultaneously all demands of travellers, thus causing an increase in travel time. Congestion is therefore closely related to the characteristics of transport infrastruc-

Francesca Medda is a tutorial fellow in the Department of Geography and Environment at the London School of Economics. E-mail: f.medda@lse.ac.uk. *Peter Nijkamp is a professor in Regional, Urban and Environmental Economics and Economic Geography, Faculty of Economics, Free University, Amsterdam. E-mail:* pnijkamp@feweb.vu.nl. *Piet Rietveld is a professor in Transport Economics, Faculty of Economics, Free University, Amsterdam. E-mail:* prietveld@feweb.vu.nl.

Geographical Analysis, Vol. 35, No. 1 (January 2003) The Ohio State University
Submitted: November 19, 2001. Revised version accepted: June 27, 2002.

ture (such as the capacity of the road) and to the characteristics of traffic flows (such as the composition of the traffic flows), and it is therefore indirectly but certainly related to the formation of the city's specific shape.

From the point of view of congestion, both the *density* and the *volume* of traffic are two fundamental traffic parameters that are able to yield better insight into urban land-use. In urban economics literature, a standard reference is Solow and Vickrey (1971), who assume in their model that the business districts in a city are uniformly distributed in the entire urban area and that each business district is an origin and a destination of traffic flows. The objective of their theoretical model is to examine the optimal allocation of urban land for transportation. Volume and density of traffic have, therefore, a prominent place in their analysis of urban form.

The aim of the present paper is to re-examine the model of Solow and Vickrey (1971) and to extend and generalise the spatial scope of the model. In particular, our attention will focus on how, in mathematical terms, transport may shape cities and thus on how the minimization of transport costs and the emerging new city shape may determine the spatial layout of the urban transport network. Before proceeding with our analysis, we will first address a number of extensions of the traditional Solow-Vickrey model in order to put our work in context in relation to previous analyses.

EXTENSIONS OF THE SOLOW-VICKREY MODEL

In this section we examine some of the most relevant examples in the literature that attempt to reformulate the structure of the Solow and Vickrey model. The models of Mills and de Ferranti (1971) and Solow and Vickrey (1971) are the first attempts in the literature to study the allocation of urban land between transport and other land uses by means of sophisticated mathematics-calculus of variations, linear programming and control theory.

Mills and de Ferranti (1971) observe that congestion costs may determine different location choices and travel patterns in relation to the standard model of Alonso (1964). This is due to the distortion of the land rent values which do then not reflect the opportunity cost of land. In fact, when deciding on their residential location, households do not consider the congestion costs imposed on other residents; they take into account only the average travel cost and thereby set in motion a chain of events, the most significant being both a greater use of transport and a population increase within the city.

The model of Solow and Vickrey (1971) analyses the case of a road network laid out in a "long narrow" rectangular city in order to minimize the aggregate transport cost incurred in the city per unit time. Solow and Vickrey (1971) consider a rectangular urban shape with the business district distributed uniformly in the city and with a given total road area. The volume of traffic is generated only by the business activities, and the cost per ton yard of moving traffic at a specific point x is assumed to be proportional to the density of traffic at x. The transport cost encompasses the congestion cost and the travel cost, i.e., the cost of travel when no other traffic is on the road. Two possible solutions arise from the model, that is, the optimal city size solution and the constricted city solution. In the former case, "the required business district can be accommodated at the least transport cost; congestion and distance costs are just in balance. The road of constant width characterizes the optimal amount of land available" (Solow and Vickrey, 1971).

In the constricted city solution, the urban area is constricted due to topographical or administrative reasons, and the road area is laid out in a rhomboid shape, that is, the road widens steadily from one boundary of the city towards the centre of the city where it reaches its maximum width and next narrows symmetrically to a zero width at the other boundary of the city. Therefore, the proportion of land allocated to trans-

portation is a decreasing and concave function of the distance from the centre of the city. By assuming that (1) the urban land is rented competitively, (2) no congestion toll is levied on road users, (3) transportation costs are absorbed at the trip origin by sellers who deliver at a uniform price, and (4) the opportunity cost of land is zero, the Solow-Vickrey model demonstrates that the rent function defined in the model "rises less sharply than it should toward the city centre because locational advantages are not fully reflected in private costs" (Solow and Vickrey, 1971). This leads to the consequent over-construction of roads, especially at the centre of the city.

Next, the Legey, Ripper and Varaiya (1973) model corroborates the results of the Solow and Vickrey model. But they distinguish between the optimal city and the market city in their model. In the optimal city, as in the Solow and Vickrey model, a central authority determines the allocation of land that minimizes social costs. In the market city, given the rent profile, the landlords maximize their profits subject to the amount of capital devoted to housing, while central authorities maximize their net benefits subject to the land and the capital devoted to transport. In the case of the market city, the Legey, Ripper, and Varaiya model demonstrates that "the market city devotes too many resources to housing at the periphery and too many resources to transportation at the centre" (Legey, Ripper, and Varaiya 1973).

The first extension of the Solow and Vickrey model where a different urban shape is examined, is the model of Kraus (1974). Kraus (1974) considers the same assumptions as made by Solow and Vickrey (1971), but by introducing a circular shaped city he extends the problem from a one-dimensional operation of the transport network to a radial and circumferential road system. Given the morphological specification of the city, Kraus (1974) mainly examines the problem of the optimal route utilization within the city. The study of the optimal allocation of land to transportation is confined to his Theorem 3, whereby Kraus (1974) states that a necessary condition for optimal land allocation between radial and circumferential roads is that the marginal rate of substitution of land for variable trip costs in radial travel and circumferential travel be equal in the production of radial and circumferential travel. The model therefore, shows that the provision of both the inner and outer ring road is optimal.

A transport constellation similar to the model of Kraus (1974), but with an alternative urban shape, is that of a square city developed in a study by Miyao (1978, 1981). Miyao (1978, 1981) considers a square shaped city, with a grid of horizontal and vertical roads. Miyao (1978, 1981), like Kraus (1974) in his model, examines the equilibrium traffic patterns in the city. Every road user selects a cost-minimizing route between points of origin and destination which are the boundaries of the smallest rectangle, which includes the points of origin and destination at its corners.

By using the same assumptions as Solow and Vickrey (1971), Miyao (1978, 1981) examines the competitive rent profile in relation to the differential transport costs. Given that the traffic volumes and densities are identical across all horizontal roads, in all points of the vertical axis that crosses at the midpoints of the horizontal roads, the marginal benefits are the same if we increase the width of the road. But the profile of the competitive rent has a maximum at the city centre and minimum at the boundary points along the vertical axis. This leads to the conclusion that "the land-use decisions based on market rents tend to create too many roads everywhere in the city at the optimum, and the tendency toward excessive road-building is stronger near the city boundaries, especially at the midpoints of the boundary lines, rather than in the city centre" (Miyao, 1978). If we consider the transport cost of vertical movement as negligible, we can derive the Solow and Vickrey one-dimensional model from the Miyao model.

The models under scrutiny in our analysis will adopt a standard definition of the volume of traffic, but they assume a specific definition of density of traffic. Standard definitions are available in the traffic engineering literature (see Daganzo, 1997). The

Solow and Vickrey model—and the subsequent extensions of it—assume that each unit of business area generates g tons of traffic per unit of time, and that $y(x)$ is the total business located at the left hand side of point x, whereas A is the total business area of the city. Given these assumptions, the total *volume* of traffic per unit of time passing the point x in both directions is $v(x) = (2g/A)y(x)\ (A-y(x))$. The *density* of traffic at x is assumed to be a function of the ratio of traffic volume and the width of the road at the horizontal coordinate x. However, if we analyse these two definitions from a dimensional point of view, we observe that the *equations of state of traffic fluidity* are not satisfied, which functionally relate density and volume of traffic to the average traffic speed. This is because Solow and Vickrey (1971) and many other authors in the field deal with the term 'density' by using a non-standard definition, even though this does not necessarily imply invalid results.

As we have shown, the Solow and Vickrey model has induced various follow-up analyses, such as the optimal route choice in a city, the optimal allocation of traffic green zones, the effect of congestion on a city, and the analysis of the competitive rent profile. From the various transport land-use interaction analyses, we focus here on the main core of the Solow and Vickrey model, namely the optimal allocation of road space in a city given by the minimization of the aggregate transport cost. In particular, our objective is to develop a generalized analytical formulation to address the optimal allocation of urban land for transportation.

URBAN SHAPES AND TRANSPORT

Solow and Vickrey (1971) assume in their model a city with a rectangular shape of length L and width W. No residential land used is considered. An area $A < WL$ is devoted to the homogeneous distribution of business firms. By assuming $w(x)$ to be the width of the road at a horizontal coordinate x, the total business area located at the left of x is given by:

$$y(x) = \int_0^x (W - w(u))du\ .$$

Given that $y(0) = 0$ and $y(L) = A$, we can now express $w(x) = W - y'(x)$. The traffic is composed by a single transport mode. Transport cost (i.e., cost per ton yard of moving traffic at x) is a function of the density of traffic at x: $f(v(x)/w(x)) = d[v(x)/w(x)]^c$, where d and c are two parameters strictly larger than zero. The transport cost is negligible in the breadth-wise direction. By assuming that a unitary business area produces g tons of traffic, the total right-to-left volume of traffic equals $v(x) = (2g/A)y(x)\ (A-y(x))$, so that the total transport cost in the entire city is:

$$\int_0^L v(x)f\left(\frac{v(x)}{w(x)}\right)dx = \left(\frac{2g}{A}\right)^{c+1} d\int_0^L \frac{y^{c+1}(A-y)^{c+1}}{(W-y')^c}dx\ . \qquad (1)$$

where:

$v(x)/w(x)$ is the density of traffic at location x;
$w(x)$ is roadway width;
c, d are constants with $c, d > 0$.

The solution of the optimal allocation of road space in the model is given by the solution of the minimization of the total transport cost as defined in (1), subject to $y(0) = 0$,

$y(L) = A$. By means of calculus of variations, Solow and Vickrey (1971) obtain then the area necessary for transport in the case of a rectangular city shape. Based on a rectangular restricted city,[1] they find that land allocated for transport is large near the centre and smaller near the city's edges. The remaining area of the city is devoted to business purposes. This outcome is obviously driven by the assumption of a rectangular form of the city (used for business and transport). In the remainder of this analysis we address the question of how to analytically express the urban shape when this assumption is dropped. Our approach is to take the shape of the business district as given and derive the corresponding land demand for transport on the basis of it. The urban shape follows as the sum of the two. The next question to be addressed is how, by assuming the configuration of the business district, i.e. the shape function of the business district, we can determine the corresponding shape of the city.

The key assumptions of the Solow-Vickrey model are also incorporated in our analysis. In particular, the urban area is defined only by both business use and road use. The areas allocated for business and transport are, respectively, the total area for business and the total area devoted for road infrastructure in the city. The sum of these two areas is equal to the total area of the city.

The width of the city is now an endogenous variable in our approach. Solow and Vickrey (1971) assume that the shape of the city is given and that all businesses are dispersed within this urban area. Let us assume here, as a contrast to the Solow-Vickrey approach, that the business district, where all business firms are located, has a rectangular shape. In order to determine the total urban area and thus the city shape, we will first determine the total area allocated for transport. We can then add up this area with the given business area.

We assume that the transport cost incurred by the business area between 0 and x is a function of the total volume of traffic passing at x from left to right; thus, we adopt this Solow-Vickrey definition of transport cost. Our objective is to define the area in the city that needs to be allocated for transport use if we assume a constant ratio of volume to capacity of traffic in the CBD.

In order to identify the city shape, we will first determine the area allocated for transport. Because we want to take into account travel flows from right to left and travel flows from left to right in the business district, we divide the business district into two equal parts along the lengthways side. In this case we obtain two equal rectangles. For any value of business area at x, $y(x)$, we have $gy(x)$ tons of goods moving to various destinations. Following the Solow and Vickrey hypothesis, we assume that the trips with destinations to the right of x are proportional to the area to the right of x in relation to the business area considered. The volume of traffic at x is next given by:

$$v(x) = g\int_0^x a\,du * \frac{\int_x^L a\,du}{\int_0^L a\,du} \qquad L \geq u \geq 0 \tag{2}$$

where:

1. We focus our attention on the case of the restricted city, since we believe it is more realistic to consider a city where its area is restricted by its physical as well as political restrictions on its area.

a is the (constant) width of the business district;
L is the length of the business district;
g are the tons of goods for the unitary business area.

The solution of the integral is: $v(x) = gax\frac{L-x}{L}$. Note that this equation is consistent with the more general formula for the entire city (which is $v(x) = (2g/A)y(x)(A-y(x))$), since in our case we have:

$$y(x) = (x / L)A$$

and

$$La = A\ .$$

The transport cost is a function of the volume of traffic divided by the function of the area allocated to transport in the city. If we assume a constant transport cost per unit distance throughout the city (denoted as R), we then obtain:

$$R \equiv T(x) = \left(g\int_0^x adu * \frac{\int_x^L adu}{\int_0^L adu} \right) * \frac{k}{w(x)} \qquad L \geq x \geq 0 \tag{3}$$

where:

$T(x)$ is the transport cost for $L \geq x \geq 0$;
k is cost per ton yard of moving traffic at x;
$w(x)$ is the function of the area allocated for transport at location x.

Given the value of $T(x)$ for all x equal to R, we can evaluate the function $w(x)$. After having calculated the functional form of $w(x)$, we can then easily integrate $w(x)$ for $L \geq x \geq 0$ and then obtain the area allocated for transport for the considered area of business district. We must remember that $w(x)$ calculated in (2) will define only *half* of the area allocated for transport, since we have divided the business district area into two equal parts. We find that the total area allocated for transport in the city has an quadratic shape (see Figure 1).

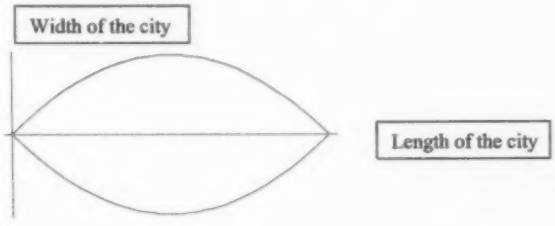

FIG. 1. Shape of the Urban Area Allocated for Transport

Such a shape was expected, because we are considering a rectangular shape of the business district, and the central point of the business district is therefore the point with the highest volume of traffic. It is easy to demonstrate in this case that the city will have a shape given by the sum of a rectangular shape and a quadratic shape. Thus, with rather simple information on the shape of the business area, we have deduced qualitative information about the urban shape.

In the next section we examine a particular structure of the business area, i.e., a polycentric structure. The structure of cities is often characterized by the combination of various centres that cluster in small urban areas in relation to the total metropolitan area. We have introduced this particular clustering urban structure as a polycentric urban structure, and in the next section we examine how to apply the methodology developed here in order to depict the shape of the polycentric city.

AN EXAMINATION OF POLYCENTRIC URBAN SHAPES

To analyse a city with a polycentric structure through the use of the methodology we have previously described, we need to know the shape of the business district. In the previous section we have assumed a rectangular shape; in this case, where we want to examine a polycentric city, we need to consider a business district having more than one CBD. In our first case, the polycentric business district is characterized by a number of business clusters of equal size. Since we need to consider the lengthways section of the business district as we have done in the preceding section, a valuable approximation of this half of the business district shape is represented by the trigonometric function of the sine.

Our consequent decision to approximate the one half of the business district shape by means of the sine function is based on two primary reasons. From an analytical point of view we obtain an elegant simplification of the business shape. Secondly, we need to remember that through the use of Fourier theory we are able to reduce any arbitrary function into an infinite series of sines and cosines; this choice may therefore be regarded as a general case when we consider more complex functions of the business district shape.

We assume that the relevant half of the business district shape of the city can be described by the following function:

$$F + \sin(x) \qquad L \geq x \geq 0 \tag{4}$$

The value of F must be strictly greater than 1 in order to have a minimum urban area between the CBDs (See Figure 2).[2]

If $F = 1$, we obtain separated urban areas and are therefore unable to examine a polycentric city. However, in this case we can study a stylised "edge city" structure. Given the shape of the business district, we can easily define the volume of traffic at x as:

$$v(x) = g\int_0^x [F + \sin(u)]du * \frac{\int_x^{5\pi} [F + \sin(u)]du}{\int_0^{5\pi} [F + \sin(u)]du} \tag{5}$$

2. In order to achieve more clarity in the exposition, the figures in this section have the same dimension but not the same scale.

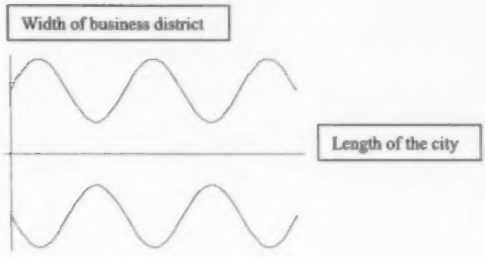

FIG. 2. Shape of the Business District
Case 1

We are assuming that the city is composed of three CBDs located at $\pi/2$, $5/2\,\pi$ and $9/2\,\pi$. In this case (as in the previous), we consider the formulation of the volume of traffic according to the definition given by Solow and Vickrey (1971).

We calculate the function of the area allocated for transport in half of the city. We are assuming a constant transport cost in the entire city. We then obtain:

$$Rw(x) = gk\int_0^x [F + \sin(u)]du * \frac{\int_x^{5\pi} [F + \sin(u)]du}{\int_0^{5\pi} [F + \sin(u)]du} \qquad 5\pi \geq x \geq 0 \qquad (6)$$

We assume $F = 2$ and $gk/T = 1$. In order to calculate the area allocated for transport in the entire city, we must sum the result in equation (6)—related to half of the city—with the symmetric result, which is the total area dedicated to transport in the city. The following figure depicts the above calculation process:

Figure 3 represents the shape of the area allocated for transport. In this case the shape of the transport area has lost the periodic pattern, which is the main characteristic of the shape of the business district. We also observe the dominant position of the central CBD in relation to the other two CBDs. By dominant position we mean that the geometrical centre of the city has the highest volume of traffic of the city. The total shape of the city, given by the sum of the business district area and the area allocated for transport, is shown in Figure 4. The shape of the city namely the sum of the

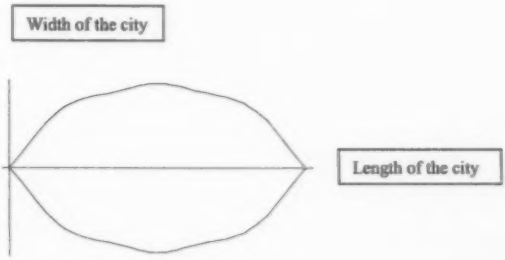

FIG. 3. Shape of the Area Allocated for Transport
Case 1

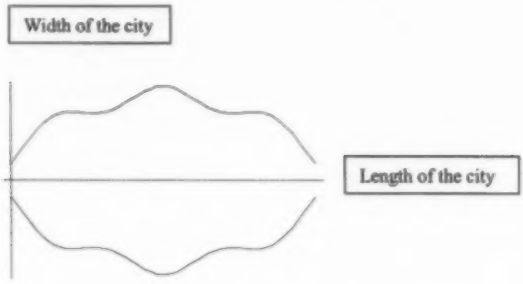

FIG. 4. Shape of a Polycentric City with a Periodic Business District Shape
Case 1

shape of the business district and the shape of the area allocated for transport, highlights the dominant role of the central CBD in contrast to the other two CBDs.

We have thus far examined a business area comprised of a number of clusters; these clusters have the same surface and shape and are therefore equal business districts. Let us now consider another type of business district shape. We examine a polycentric urban structure, but in this case we imagine that the different CBDs have different surfaces and forms; i.e., we have a large business area followed by subsequent smaller ones. The lengthways half of the business district shape is represented as follows:

$$e^{\frac{-x}{b}}[F+\sin(x)] \qquad L \geq x \geq 0 \tag{7}$$

where F is the same as in equation (5) and b is a constant that must be greater than 1 and that in our case is arbitrarily set equal to $2F$. To describe the shape of the business district we use a periodic function, but in this case we introduce a damping effect in order to obtain the chosen business district shape (see Figure 5).

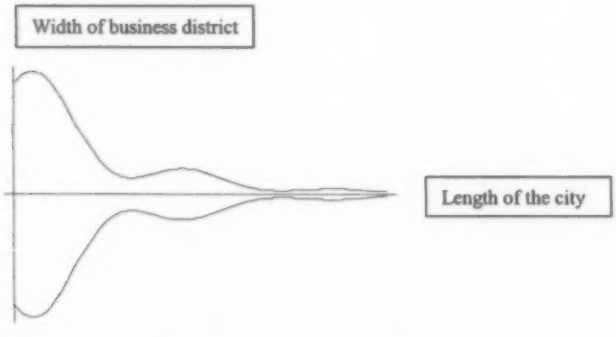

FIG. 5. Shape of the Business District
Case 2

The volume of traffic is given by:

$$v(x) = \int_0^x e^{\frac{-u}{b}}[F + \sin(u)]du * \frac{\int_x^{5\pi} e^{\frac{-u}{b}}[F + \sin(u)]du}{\int_0^{5\pi} e^{\frac{-u}{b}}[F + \sin(u)]du} \qquad 5\pi \geq x \geq 0 \qquad (8)$$

We are assuming here a constant transport cost. We can now evaluate the function $w(x)$ that indicates the area allocated for transport in the one half of the city:

$$Rw(x) = gk\int_0^x e^{\frac{-u}{b}}[F + \sin(u)]du * \frac{\int_x^{5\pi} e^{\frac{-u}{b}}[F + \sin(u)]du}{\int_0^{5\pi} e^{\frac{-u}{b}}[F + \sin(u)]du} \qquad 5\pi \geq x \geq 0 \quad (9)$$

As in the previous case, by assuming that $F = 2$ and $gk/T = 1$, we are able to depict the shape of the area allocated for transport (see Figure 6).

The shape of the transport area in this second case results from the dominant role of the main CBD: the majority of the area allocated for transport is, in fact, in the surrounding area of the largest CBD. The implication is that the dominant CBD is the main origin and destination of urban trips. Our next step is to calculate the entire area and thus to obtain the shape of the city by summing up the area allocated for business uses and for transportation. We derive then the following city shape (see Figure 7).

The shape of this city with a business district shape thus defined may, for example, correspond to the shape of port cities, in which the main CBD is nearest to the port, and the city grows outward from this central point (like, e.g., in Chicago, U.S.A.).

The method developed here appears thus to be a suitable tool for analysing the approximation of urban shapes due to the effects of traffic flows. Our analysis highlights the paramount role of road traffic in the growth and development of urban structures, and particularly in the spatial definition of the city.

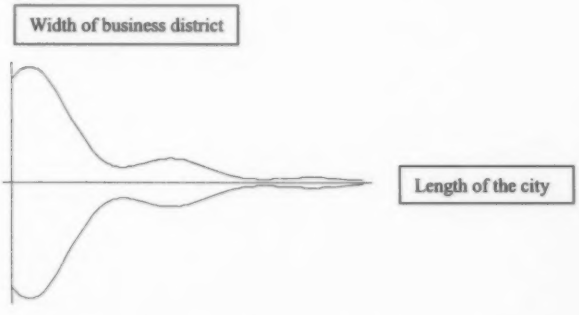

FIG. 6. Shape of the Area Allocated for Transport
Case 2

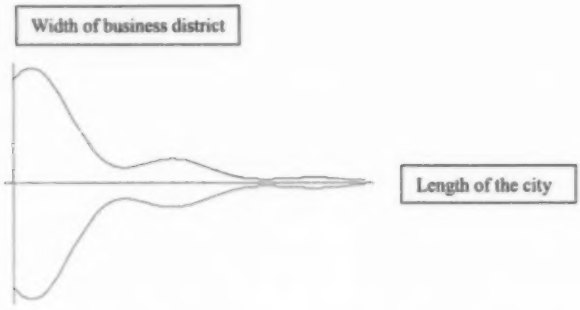

FIG. 7. Shape of a Polycentric City with a Periodic Damped Business District Shape Case 2

CONCLUSION

The model of Solow and Vickrey (1971) and subsequent extensions have been used here in order to address the problem of the allocation of urban land to transport. The analysis departs from the Solow-Vickrey foundation to investigate the impact of traffic flows on the urban shape definition. Particular attention has been given to two specific cases of a polycentric urban structure.

The two cases emerge from the necessity to discover the spatial elements that may have generated the city shape. As a first step we have revised the Solow-Vickrey approach, by assuming as given, the shape of the business district and then we determined the shape and area of the urban land use allocated for transport. The sum of the business area and transport area allows us to define the morphology of the city. The objective of the model has been to determine the mathematical formulation for defining the shape of the city when we assume a polycentric shape of the business district. To ascertain the shape of the business district, we approximated two possible business districts. The two different shapes were chosen to investigate how different geometrical characteristics influence the urban shape definition. In the first case we have considered a polycentric structure with three CBDs, each having the same area and thus being of equal importance. When defining the urban shape resulting from this business district shape, we noticed how the geometrical relation between the CBDs leads to a dominant role of the central CBD. In other words, the central CBD is not only an origin and destination of traffic, but it is also a location through which all traffic passes.

In the second case we have considered the structure of the business district where the CBDs were positional in a dominant hierarchical structure. In this case we have seen that the geometrical relationship plays a more subdued role in relation to the hierarchical relationship among the CBDs; i.e., the CBD with a greater business area requires the highest share of land for transport.

Although based upon a possible approximation of the shape of the business district, the methodology discussed here provides a useful analytical tool for evaluating the impact of traffic flows, and suggests also a possible means for alleviating detrimental impacts of spatial externalities. Above all, we have highlighted in our analysis the indirect crucial role that transport plays in urban morphology formation.

LITERATURE CITED

Alonso, W. (1964) *Location and Land Use*. Cambridge: Harvard University Press.

Berry, B.J.L. and H. Kim (1993) Challenges to the monocentric model. *Geographical Analysis* 25, 1–4.

Cervero, R., and K-L. Wu (1997) Polycentrism, commuting, and residential location in the San Francisco Bay area. *Environment and Planning A* 29, 865–886.

Daganzo, C.F. (1997) *Fundamentals of Transportation and Traffic Operations*. New York: Pergamon.

Kraus, M. (1974) Land use in a circular city. *Journal of Economic Theory* 8, 440–457.

Ladd, H.F., and W. Wheaton (1991) Causes and consequences of the changing urban form: introduction. *Regional Science and Urban Economics* 21, 157–162.

Legey L., M. Ripper, and P. Varaiya (1973) Effects of congestion on the shape of a city. *Journal of Economic Theory* 6, 162–179.

Martinez, F., and C. Araya (2000) Transport and land-use benefits under location externalities. *Environment and Planning A* 32, 1611–1624.

Mills, E.S., and D.M. de Ferranti (1971) Market choice and optimum city size. *American Economic Review* 61, 340–345.

Miyao, T. (1978) A note on land use in a square city. *Regional Science and Urban Economics* 8, 371–379.

Miyao, T. (1981) *Dynamic Analysis of the Urban Economy*. New York: Academic Press.

Peeters, D., J-F. Thisse, and I. Thomas (1998) Transportation networks and the location of human activities. *Geographical Analysis* 30, 325–347.

Solow, R.M., and W.S. Vickrey (1971) Land use in a long narrow city. *Journal of Economic Theory* 3, 430–447.

Wang, F. (2000) Modelling commuting patterns in Chicago in a GIS environment. *Professional Geographer* 52, 120–133.

Wheaton, W.C. (1998) Land use and density in cities with congestion. *Journal of Urban Economics* 43, 258–272.

Peter M. Atkinson
Sally E. German
David A. Sear
Michael J. Clark

Exploring the Relations Between Riverbank Erosion and Geomorphological Controls Using Geographically Weighted Logistic Regression

The relations between riverbank erosion and geomorphological variables that are thought to control or influence erosion are commonly modelled using regression. For a given river, a single regression model might be fitted to data on erosion and its geomorphological controls obtained along the river's length. However, it is likely that the influence of some variables may vary with geographical location (i.e., distance upstream). For this reason, the spatially stationary regression model should be replaced with a non-stationary equivalent. Geographically weighted regression (GWR) is a suitable choice. In this paper, GWR is extended to predict the binary presence or absence of erosion via the logistic model. This extended model was applied to data obtained from historical archives and a spatially intensive field survey of a length of 42 km of the Afon Dyfi in West Wales. The model parameters and the residual deviance of the model varied greatly with distance upstream. The practical implication of the result is that different management practices should be implemented at different locations along the river. Thus, the approach presented allowed inference of spatially varying management practice as a consequence of spatially varying geomorphological process.

The authors thank the Environment Agency, Welsh Region and the Department of Geography, University of Southampton for research funding to Sally German. The GeoData Institute, in particular, Chris Hill and James Feaver are thanked for general support and for the provision of computing facilities. Finally, we thank the anonymous referees for their constructive comments.

Peter M. Atkinson is a professor in the Department of Geography at the University of Southampton. E-mail: P.M.Atkinson@soton.ac.uk. *Sally E. German is a research scientist at the Geodata Institute at the University of Southampton. Email:* seg@geodata.soton.ac.uk. *David A. Sear is a senior lecturer in the Department of Geography at the University of Southampton. Email:* D.Sear@soton.ac.uk. *Michael J. Clark is a professor in the Department of Geography at the University of Southampton. Email:* M.J.Clark@soton.ac.uk.

Geographical Analysis, Vol. 35, No. 1 (January 2003) The Ohio State University
Submitted: July 1, 2001. Revised version accepted: July 10, 2002.

INTRODUCTION

The evolution of river planform is an important area of research in fluvial geomorphology, and is of significance to river managers and owners of riparian land. Bank erosion, a key process of planform change, continues to be a focus for process and applied research. The erosion of river banks not only causes change in channel planform and cross-section but is a key process whereby floodplain sediments are remobilised to become a part of the basin sediment yield (Bull 1997).

The literature associated with channel planform change is now extensive, with research largely focusing at two scales: (i) mesoscale analysis of channel planform and bank retreat rates over periods of up to 200 years (Gurnell 1997) and (ii) more detailed site-specific studies of the processes of bank retreat (Hooke 1995; Lawler, Thorne, and Hooke 1997; Bull 1997). Relatively few studies attempt to link such different scales of analysis, and even fewer attempt this at the basin scale. Notable exceptions include the work of Lawler (1992) who identified theoretical process domains for bank erosion, and Abernathy and Rutherford (1998) who attempted a similar analysis including the additional influence of vegetation. However, even these studies involve fragmentary information and make no explicit attempt to link historical channel activity with the full suite of contemporary controls on the physical processes of bank erosion.

Recent geomorphological investigations into the processes of bank erosion have been undertaken through the use of erosion pins and PEEPS (Lawler 1993; Lawler, Thorne, and Hooke 1997), investigations into bank stability (Thorne 1982; Simon and Hupp 1992; Simon and Darby 1995), and the use of bank reconnaissance methods (Thorne, Allen, and Simon 1996). However, these investigations tend to focus on particular banks and specific erosion issues, and the detailed data sets required preclude investigation of the entire catchment. Further, the multivariate nature of the controls and the variability of these in space can prevent identification of the actual processes of bank erosion. In particular, geomorphologists are frequently challenged to explore relationships that may vary in their nature from place to place (they require spatially non-stationary models) and that involve categorical data.

Bank erosion is a significant problem for many people who derive income from, or live on, land next to rivers. Bank erosion management has typically been site-specific (tending to treat the problem where it arises) and simplistic (using techniques centred on the assumption that the erosion is caused by fluvial scour) (Sear, Newson, Brookes 1995; Thorne 1997). Recent application of geomorphological theory to bank erosion management has focused on, first, identifying the cause of erosion and, second, developing a more sustainable bank management plan in partnership with engineers and stakeholders. At the heart of this philosophy is the ability to identify the dominant processes and the controls on these processes operating at the bank scale, whilst recognising that the causes may lie at the reach or catchment scale. To date, this process has been largely field-based and qualitative, with interpretation of the evidence relying on the professional capability of those involved. What is needed is a more objective framework for integrating and assessing different types of data across a broad range of spatial scales.

In this paper, new survey and statistical techniques are presented that can be applied at the basin scale, and that integrate spatially distributed historical data on the timing of river channel planform change with spatially distributed data on the properties controlling bank failure. A method for the analysis of spatial variation in the relations between bank erosion and its controlling factors is presented that incorporates non-stationary parameters and categorical data. This method is termed geographically weighted logistic regression (GWLR). The objective of this paper is to apply

GWLR to model the spatial variation in the relation between erosion (presence or absence) and several controlling variables for the Afon Dyfi in West Wales.

THEORY

Statistical modelling has been applied extensively to fluvial systems (see Rhoads 1992). In this section, the basic linear regression model and its extension to (i) spatially varying parameters (GWR) (Brunsdon, Fotheringham, and Charlton 1996) and (ii) prediction of a binary response variable (logistic regression) are described.

Linear Regression

Multiple linear regression is commonly applied in geomorphology and the steps in modelling and the assumptions involved are well understood. Therefore, only the basic linear regression model is given here,

$$y_i = \beta_0 + \sum_{k=1}^{m-1} \beta_k x_{ik} + \varepsilon_i \tag{3}$$

predicting the target variable y_i at location i as a function of m parameters β_0 and β_k ($k = 1,2,\ldots, m-1$) and $m-1$ explanatory variables x_{ik}. The ε_is are independent normally distributed error terms with zero means. Usually, the least squares method is used to estimate the β_ks. This is most readily expressed using matrix notation as

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} \tag{4}$$

where $\hat{\boldsymbol{\beta}}$ is a single column vector of coefficients, the independent variables fill the columns of $\mathbf{x}$, the dependent variable fills the single column of $\mathbf{y}$ and t indicates the transpose of a matrix (Brunsdon, Fotheringham, and Charlton 1996). To estimate β_0 $\mathbf{x}$ must contain a first column of 1s.

Equation 3 is implicitly a *stationary* model of the relations between $\mathbf{y}$ and $\mathbf{x}$. A stationary model is one in which the parameters of the model do not change with geographical location. Thus, it is only appropriate where the relations may reasonably be modelled as invariant over space. Where the relations may be expected to change with location it is necessary to apply a non-stationary model: one that allows some freedom in the parameters of the model as one moves from place to place. Brunsdon, Fotheringham, and Charlton (1996) developed such a non-stationary regression model and termed it geographically weighted regression (GWR). A similar model was applied in relation to fluvial systems by Rhoads (1991). GWR offers a basis upon which to tackle some of the issues mentioned in the introduction and represents a significant potential enhancement over conventional regression approaches.

Geographically Weighted Regression

GWR extends the traditional regression model of Equation 1 to the non-stationary case by allowing the $\boldsymbol{\beta}$ coefficients to vary with location i,

$$y_i = \beta_{i0} + \sum_{k=1}^{m-1} \beta_{ik} x_{ik} + \varepsilon_i \tag{5}$$

where β_{ik} is the value of the k^{th} parameter at location i. It is worth noting that Equation 3 is a special case of Equation 5 where the parameters do not vary across space.

A problem with Equation 5 is that there is only one observation at i for each of y_i and the $m-1$ x_{ik}. To provide more observations with which to build a regression model for each i, Brunsdon, Fotheringham, and Charlton (1996) chose to consider the relation between **y** and **x** within a region defined as geographically proximate to each i, while acknowledging that such an approach involves approximation.

The weighted least squares approach to calibrating regression models provides a means by which to vary the influence of observations on the resulting model. Specifically, in weighted least squares a weight w_i is applied to each squared difference before minimization so that the inaccuracy of some predictions carries more weight than that of others. If **w** is the diagonal matrix of w_is then the estimated coefficients satisfy:

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{w}\mathbf{x})^{-1}\mathbf{x}'\mathbf{w}\mathbf{y} \tag{6}$$

Since the weight **w** varies with location j, such that a different calibration exists for each j, it is more appropriate to write Equation 6 more explicitly as:

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{w}(\mathbf{j})\mathbf{x})^{-1}\mathbf{x}'\mathbf{w}(\mathbf{j})\mathbf{y} \tag{7}$$

Note that location has now been referenced by j since the location at which a model is calibrated need not equal the location of available data i.

Brunsdon, Fotheringham, and Charlton (1996) consider a range of possible weighting functions for use in Equation 7. These include a simple step function,

$$\begin{aligned} w_{ij} &= 1 \qquad \text{if} \quad d_{ij} < d_j \\ w_{ij} &= 0 \qquad \text{otherwise} \end{aligned} \tag{8}$$

where d_{ij} is the distance between the location i of any observation and the location j at which the model is being calibrated. This weighting function has the advantage that the number of data used in the regression model is limited resulting in increased speed of computation. However, it has the disadvantage that its discontinuous nature can result in artefacts in the spatial variation in the regression coefficients. An alternative is to specify w_{ij} as a *continuous* function of d_{ij}. The exponential function is an obvious choice,

$$w_{ij} = \exp\left(\frac{-d_{ij}}{a}\right) \tag{9}$$

where a is a non-linear parameter. As a compromise between the computational saving of Equation 8 and the continuity of Equation 9, Brunsdon, Fotheringham, and Charlton (1996) suggest using the bisquare function:

$$\begin{aligned} w_{ij} &= [1-d_{ij}^2/d_j^2]^2 \qquad \text{if} \quad d_{ij} < d_j \\ w_{ij} &= 0 \qquad\qquad\qquad\quad \text{otherwise} \end{aligned} \tag{10}$$

The essential idea of GWR, whatever weighting function is chosen, is to give more weight to observations close to the location j at which the regression model is desired than further away. In this sense, the method draws strong parallels with geostatistical techniques for local prediction (Matheron 1965, 1971; Goovaerts 1997).

Logistic Regression

Generalised linear modelling (GLM) (Aitkin et al. 1989; Collett 1991) allows the formation of a (multivariate) regression relation between a response variable and explanatory variable(s). All variables may be either continuous or categorical, or any combination of both types. GLM has at least two advantages over traditional regression. First, GLM allows one to include variables that are categorical and many, if not most, variables in geomorphology are categorical. Second, parameters do not need to be normally distributed about their mean and no assumptions are made about their error distributions. Thus, most of the limitations of traditional regression are removed. Examples of this type of modelling in geomorphology include Wang and Unwin (1992), Uno, Sugii, and Hayashi. (1994), and Atkinson and Massari (1998).

In this paper, our interest is the presence or absence of erosion. The most natural regression model for this type of binary response (with 1 = "erosion" and 0 = "no erosion") is the logistic where the usual linear predictor is augmented by a logistic link function. The full logistic regression model is thus,

$$\log\left(\frac{p_i}{1-p_i}\right) = \gamma_i \tag{11}$$

or equivalently,

$$p_i = \frac{\exp(\gamma_i)}{1+\exp(\gamma_i)} \tag{12}$$

where p_i is the probability of an event occurring associated with a given location i and γ_i is the linear model (equation 3). The linear model formed is then a logistic regression of success or failure of a given binary variable (for example, presence or absence of erosion) on the explanatory variables. Useful introductions to GLM are given in Baker and Nelder (1978), Healy (1988), Aitkin et al. (1989), Dobson (1990), and Collett (1991).

The model deviance D for binary data in a GLM may be expressed as:

$$D = -2\sum_{i=1}^{n} p_i \operatorname{logit}(p_i) + \log(1-p_i) \tag{13}$$

This model deviance cannot be used directly as an indicator of goodness-of-fit since one cannot compare the fitted probabilities and the binary observations (Collett 1991). However, since the difference in deviance for a given difference in degrees of freedom between separate models can be approximated to have a Chi-squared distribution, the reduction in model deviance can be used as an indicator of the significance of each term added to the model (Baker and Nelder 1978). Therefore, terms are normally added to the model in a stepwise fashion, checking the significance of each term at each step by finding the probability for a given reduction in model deviance and number of degrees of freedom. Only those terms that are significant at a given confidence level (for example, 95%) should be retained. Once a term has been checked as significant, each parameter estimate should be tested for significance against its standard error using the t-value (estimate/s.e.). At the 95% significance level a parameter estimate is significant when it is greater than approximately twice the standard error (Atkinson and Massari 1998).

A potential problem exists for categorical variables that are divided into binary terms because only certain of these terms may be significant while others may be insignificant. It is common practice to aggregate such categorical variables into fewer, coarser classes to reduce the problem and also to simplify the resulting GLM. Interpreting the resulting GLM is quite straightforward. For categorical variables, a positive value implies an above average correlation and a negative value implies a below average correlation. For continuous variables, a positive value implies a direct correlation and a negative value implies an inverse correlation, while the magnitude of the value gives the strength of the correlation.

In this paper, GWR and GLM (specifically the logistic model) are combined to form a geographically weighted logistic regression (GWLR). This combination of a spatial model with logistic regression has similarities with, but is conceptually different to, autologistic regression (Augustin, Mugglestone, and Buckland 1996). The GWLR will be used to explore the dependence on geographic location of the relation between erosion and geomorphological controls for the Afon Dyfi, Wales.

FIELD SITE AND SURVEY DATA

The Afon Dyfi

The Afon Dyfi is a gravel-bed river situated in North Wales, running partly through Snowdonia National Park. The river drains from the Aran Mountains and enters the sea through an estuary of international importance (designated as a Ramsar site and a Biosphere Reserve). The river and its catchment have a glacial legacy and are composed of two parts. The upper reaches flow through bedrock, and there is very little floodplain due to glacial terraces and valley walls. The middle and lower reaches run through alluvial and glacial deposits (Jones 1995). Here, wide floodplains and grazing land take over from bedrock gorges and wooded riverbanks, although several terraces, both glacial and fluvial, are still present in the middle reaches that confine channel movement.

Riparian landowners have identified several bank erosion issues within the Afon Dyfi that relate primarily to loss of valley-bottom farmland. In response to a series of requests for bank protection, the local river management agency initiated a study to identify a more rational, catchment-scale management strategy that accounted for the processes controlling bank erosion on the Afon Dyfi. To determine the controls and influences on the processes of erosion, large-area (both temporal, or historical, and spatial) investigations were undertaken (German 2000). The "historical" investigation provided data on the long-term stability of the planform and the "spatial" investigation identified and measured properties that defined the fluvial environment at the time of the investigation. All the data sets were contiguous (spatially) for left and right bank, from the mouth of the river to its source.

Historical Data

A historical stability index was constructed by digitising historical maps and aerial photographs from 1890 to 1992. These planforms were imported into the SPANS geographical information system (GIS) and annotated with the age of the adjacent floodplain and the number of times the channel had remained in its present position.

The age of the floodplain adjacent to the present bank line (1998) provided information on the age of the present bank material. The banks were graded according to the age of the most recent channel as shown in Table 1. The number of times the channel had been present in the same position was determined by GIS overlay and was between one and five times. From this information a historical stability index was created and each individual bank was assigned a number (between 1 and 50), as follows:

TABLE 1
The Allocation of Scores to Bank Lines of Different Dates

Date	No.
Before 1890	10
1890	8
1902	6
1948	4
1980	2

$$\text{Historical stability} = \text{age of the adjacent floodplain} \times \text{number of times channel has been in the same position} \quad (14)$$

The larger the number, the more historically stable was the bank. In addition, the width of the planform movement (swath width) was measured every 100 m to determine the magnitude of the channel change.

Field Survey

Field data were acquired during the 1998 field season using a continuous walk-through survey. Data were acquired on several variables including the location of current erosion, vegetation, and bank material. Both banks were surveyed from mouth to source. In the field, banks were classified according to the active erosion process acting on the bank (for example, overhang, cantilever, or slab failure). This information was annotated on a base map and subsequently attributed to digitised vectors within a GIS package. The same process of data acquisition was used for vegetation and bank material. To determine the relationship between the factors controlling bank erosion and the physical locations of this erosion, it was considered appropriate to reclassify the detailed erosion types into the presence or absence of erosion. Vegetation data included type, location on the bank, density, and age. All of these properties provided information related to the stability of the bank on which the vegetation grows. These data were used to create a vegetation stability index, and individual banks were allocated a number (between 1 and 97), as follows:

$$\begin{aligned} \textit{Vegetation stability} = {} & [(\textit{woody age} \times 3) \\ & + (\textit{woody density} \times 3) \\ & + (\textit{woody location} \times 3)] \\ & + [(\textit{herbaceous age} \times 2) \\ & + (\textit{herbaceous density} \times 5) \\ & + (\textit{herbaceous location} \times 3)] \\ & + [(\textit{grass density} \times 4) \\ & + (\textit{grass location} \times 3)] \end{aligned} \quad (15)$$

Equation 2 defines the weights given to different types of vegetation. The larger the number, the greater the contribution to bank stability. Woody vegetation has the same weight for each variable (age, density, and location), whereas for herbaceous vegetation the weights differ reflecting the view that age is a less important influence on bank stability than density (Thorne 1990). Age has two categories (young and mature): the older the vegetation the more stable the bank. Density has three categories (sparse, average, and dense): the denser the vegetation the more stable the bank. Location has three categories (top, middle, and bottom of the bank in any combination).

Each of the three locations has the same weighting: the more locations that the vegetation occupies the larger the score. Table 2 describes the scores assigned to each variable in Equation 2.

Bank material was classified into five categories by field observation. These were (i) cohesive, (ii) composite (30% gravel), (iii) composite (60% gravel), (iv) non-cohesive (100% gravel), and (v) bedrock. These data were allocated to every bank within the river.

Other variables used in this research were stream power (Bagnold 1966) and the location of meander impact points on the bank lines (Thorne 1997). Direct measurement of stream power was not possible for the whole river due to the absence of survey data. Instead, an approach was adopted whereby the cross-sectional area every 100 m down river was predicted from the bank height and map-derived channel widths. Channel gradient was predicted from a 1:25000 map and bankfull discharge was calculated using the regime equations of Wharton (1995). Controls on these predictions were available at five points along the channel where direct survey had been made, or where a gauging station existed. Typically, the values obtained under-predicted the locally derived values of stream power by 20%.

Locations of meander impact points were assigned to the bankline within the GIS as 0 = no meander impact point and 1 = meander impact point. Each meander impact point was spatially distributed, that is, represented a sequence of points along the river.

Data Organisation

Data associated with each bank were sorted by location. The sequence of locations was used to determine the mid-point of each individual length for each bank. The histograms for each of the five variables are shown in Figure 1 (left bank) and Figure 2 (right bank). Each variable is also plotted against position upstream in Figure 3 (left bank) and Figure 4 (right bank).

The data were entered into a GIS to allow visualisation of the individual spatial variables. However, the number of variables available and the length of the river (over 42 km) made it difficult to view the different relations between control and effect. In addition, for management purposes, it is important to determine any differences in these controls at a local level, as this information can provide suitable guidance on the type of bank protection that may need to be adopted locally. Therefore, a methodology was required that could deal with a large data set and quantify the relations between contemporary erosion and several variables at the local scale.

METHODOLOGICAL CHOICES

This section describes several of the choices made in applying GWLR to model the spatial variation in the relation between erosion (presence or absence) and geomorphological variables for the Afon Dyfi in West Wales.

TABLE 2
Scoring System Used to Construct the Vegetation Index

	(a) woody vegetation		(b) herbaceous vegetation		(c) grass	
	Score	Weight	Score	Weight	Score	Weight
Age	1 2	3	1 2	2		
Density	1 2 5	3	1 2 5	5	1 2 5	4
Location	1 2 3	3	1 2 3	3	1 2 3	3

NOTE: A score is assigned per variable (e.g., young age = score of 1) and all weighted scores are then summed to construct the vegetation index.

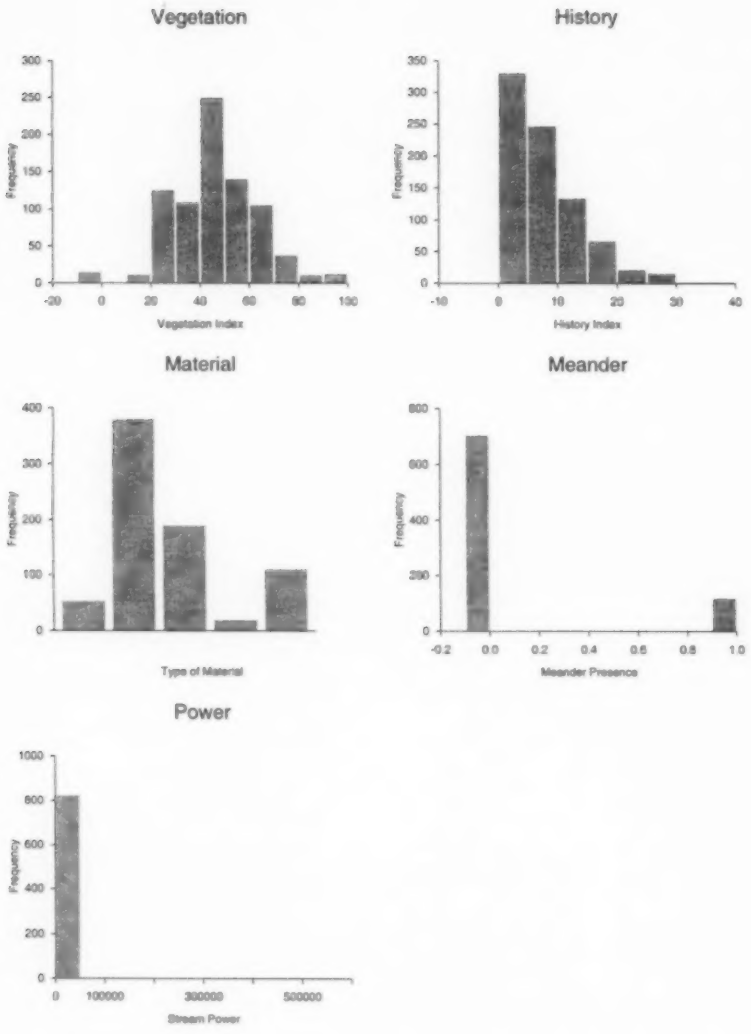

FIG. 1. Histograms of Vegetation, History, Material, Meander and Power for the Left Bank. The variables are defined in the text.

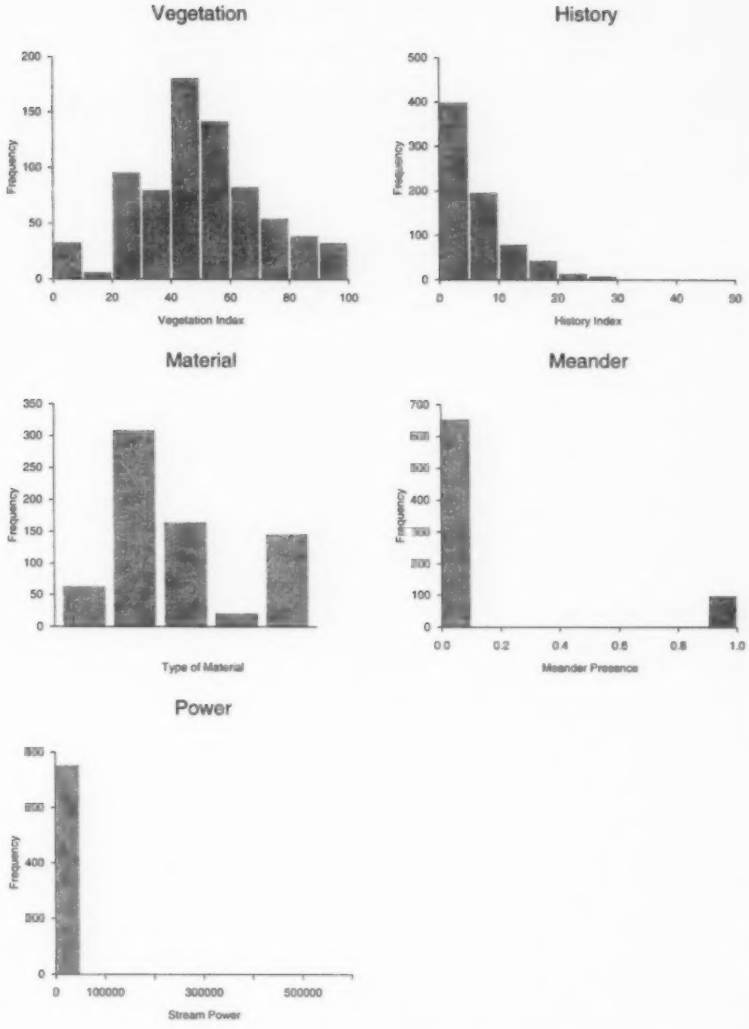

FIG. 2. Histograms of Vegetation, History, Material, Meander and Power for the Right Bank. The variables are defined in the text.

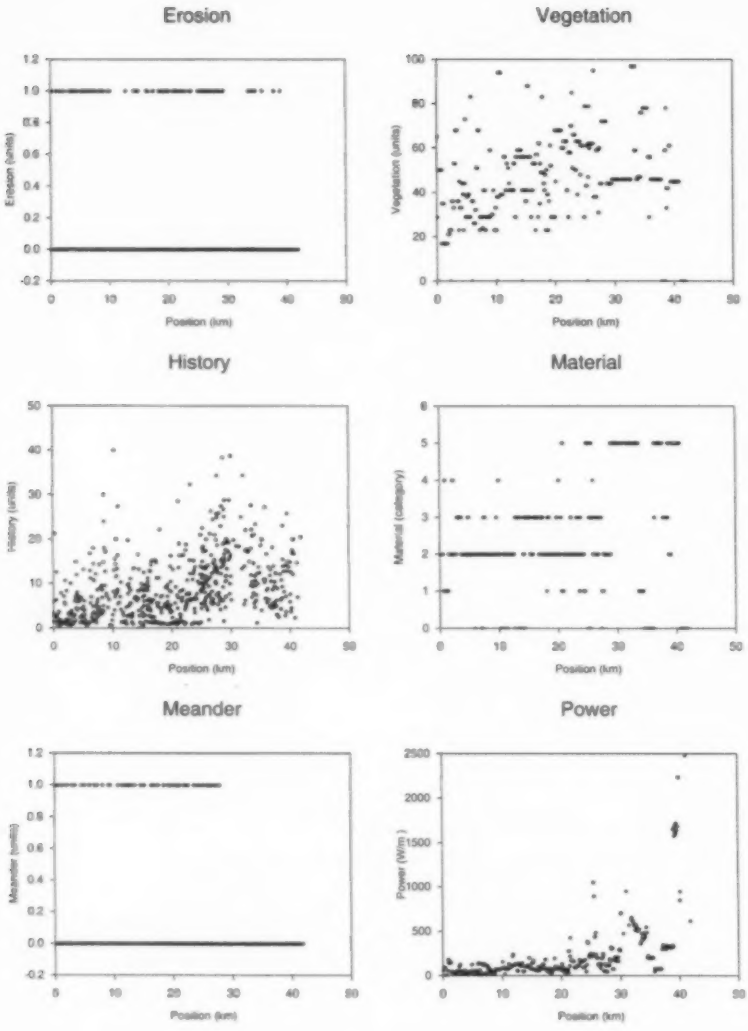

FIG. 3. Erosion, Vegetation, History, Material, Meander and Power Plotted Against Distance Upstream for the Left Bank.

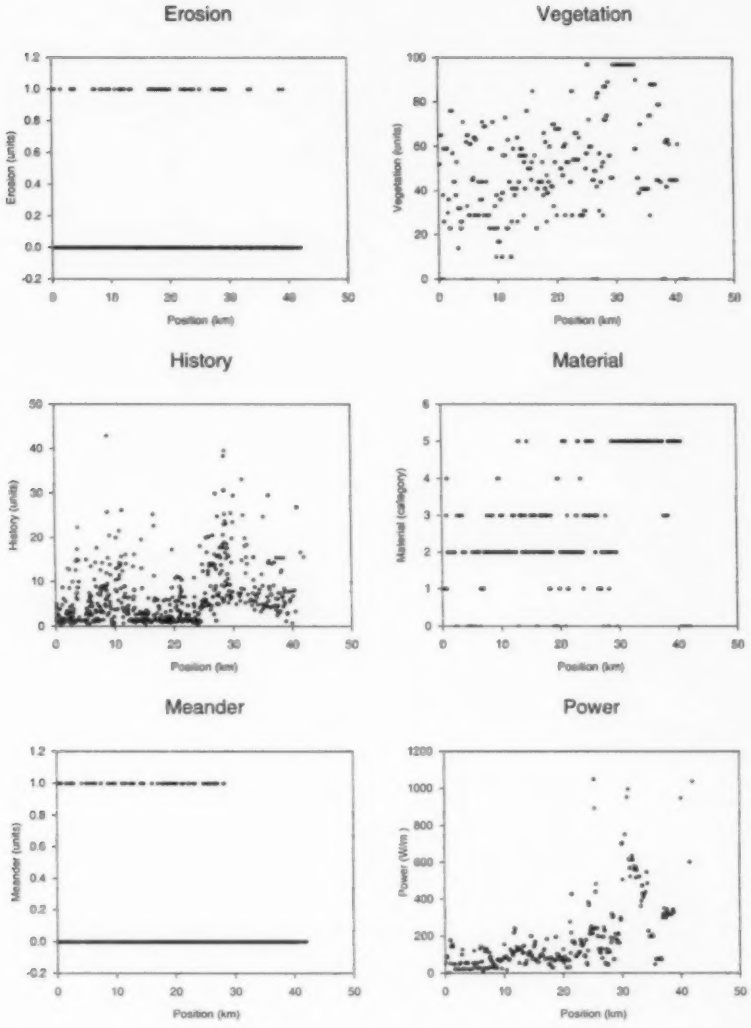

FIG. 4. Erosion, Vegetation, History, Material, Meander and Power Plotted Against Distance Upstream for the Right Bank.

Univariate Modelling

In this paper, the multivariate GWLR was not fitted. Rather, univariate GWLR models were fitted to each explanatory variable in turn. The basic rationale for this choice was that the main objective of the paper was not prediction (using the geomorphological data to determine with greatest precision the level of erosion likely), but rather to explore the relations between erosion and several individual geomorphological variables. Multivariate regression is such that meaningful relations with certain explanatory variables can become downweighted and potentially omitted from the multivariate model as other correlated variables enter the model first. To avoid this situation, variables must, at least initially, be treated individually. Additional benefits of univariate modelling are that the procedure is kept simple, the models are easier to interpret, and there are no interaction terms. However, one should remember that univariate models do not involve independent effects (i.e., controlling for the effects of other explanatory variables). Thus, the fitted parameters in a set of univariate models may be very different to those in the equivalent multivariate model.

Choice of Weighting Function

The exponential weighting function (Equation 10) with $a = 1$ km was chosen for this research because the experience of the authors in geostatistics suggests that for physical properties in geomorphology and related subjects the exponential model often describes well the autocorrelation structure. In geostatistics, and specifically the linear prediction technique known as (co)kriging, interpolation weights are chosen by reference to a model fitted to the experimental autocorrelation function (or similar) and the exponential model often provides a good fit to the experimental plot. Thus, the exponential model represents a reasonable *a priori* choice. The Gaussian model (one of several models used by Fotheringham, Brunsdon, and Charlton 2000) is, by comparison, rarely used in geostatistics and describes only the autocorrelation structure of variables that are locally very smooth (for example, digital elevation data).

Choice of Number of Data

Initially the exponential model was used to weight *all* 830 (or 758) data along the left (or right) banks of the river. However, the computational burden when running the model-fitting procedure in Splus was prohibitive, each variable for each bank taking around one hour to run. Therefore, the number of data used in the regression modelling was limited to the nearest 101 points, that is, 50 points either side of the central point i. This number was chosen as a compromise between having many data for statistical regression and few data distant from i. The computational time required for model fitting was reduced to around one-fifth of that required when using all data.

A disadvantage of using a limited number of data was that when combined with the exponential model the weights do not decrease to zero at the most distant point (as for the bisquare function), but rather decrease to a minimum. This may lead to artefacts in the maps of regression coefficients as described earlier. However, we chose this strategy because (i) the exponential model is commonly fitted to autocorrelation functions in physical geography and (ii) we wanted to fix the number of data used in modelling at each location i, (except for the ends of the river course where the number decreased). The bisquare function would involve different numbers of data being used at each location i. A comparison between the results using all data and the results using only 101 data was made and the differences were only slight.

ANALYSIS

The GWLR model was applied to each of the five explanatory variables, and for each of the left and right banks. This generated two sets of results. The first is the usual slope coefficient plotted against position. The second is a plot of {(null deviance-residual deviance)/null deviance}, here called a deviance ratio. This was provided to give some measure of the goodness-of-fit of the model at each location. The results for each of the explanatory variables are described below.

Vegetation

The model coefficients for vegetation are plotted against position in Figures 5a (left bank) and 6a (right bank), while the deviance ratio is plotted against position in Figures 7 (left bank) and 8 (right bank). There are several interesting features to note. Overall, there is a negative relation between vegetation and erosion. That is, the greater the index of vegetation the less erosion; the implication is that vegetation makes banks more stable (or, in principle, that more vegetation grows on banks that are stable) (Thorne 1990; Millar 2000). However, the nature of the relation between vegetation and erosion varies along the river channel. In particular, vegetation appears to have a greater influence on erosion for the downstream reaches (less than 20 km upstream on the plots). The influence seems to peak at around 10 km and 15–18 km upstream for both the left and right banks. This again is what might be expected since the downstream sections of the river run through alluvial deposits where the presence of additional protection from fluvial scour or increased stability of the bank materials would be more pronounced. It is worth noting that around 13 km upstream there is no relation between vegetation and erosion (for both banks) suggesting that either there is no erosion and no vegetation as bank protection is present in this area or the controlling variables do not explain erosion in this reach.

There is a large peak in the influence of vegetation on erosion at around 30 km upstream, especially for the right bank. For the upper reaches of the river, the channel is bedrock, but large glacial terraces overlain on the bedrock adjacent to the channel provide sufficient material for vegetation to influence erosion.

Not only does the strength of the relation vary with position, but also there is a direct relation between the two variables around 20 km upstream. At this point, a high vegetation index appears to correspond to large amounts of erosion. This unexpected result appears at a position where erosion is dominated by scour of the upper banks arising from the turbulence developed around widely spaced trees.

History

The most notable historical feature is the small correlation with erosion (Figures 7 and 8). Generally, one might expect banks that are eroding now to have been eroding in the past. However, contemporary erosion does not necessarily contribute greatly to historical lateral change in river planform. Thus, banks that are eroding at present are not necessarily related to planform change over the last 108 years. There are two main sets of positions at which the relation between history and erosion is inverse as expected: one at 10 km and 38 km upstream (left bank) (Figure 5) and one at about 23 km upstream (right bank) (Figure 6). At 10 km and 38 km (left bank) we expect positive erosion and a low historical index value (implying less stability). At 23 km (right bank) we expect zero erosion and a high historical index value (implying greater stability). At around 38 km upstream (right bank) there appears to be an exception to the generally inverse relation between history and erosion. We currently have no explanation for this exception, but it is worth noting that the magnitude of this effect is very small (Figure 8).

Material

The deviance ratio for material (Figure 7, left bank and Figure 8, right bank) shows that, overall, variation in material has the largest effect on erosion along the Afon Dyfi. However, as for vegetation, the influence is not constant along the river. The largest peak occurs at around 30 km, with a second peak at around 20 km (both left and right banks). These positions correspond to changes from composite banks with 30% gravel and 60% gravel to bedrock (30 km) and from composite banks with 30% gravel to composite banks with 60% gravel (20 km). There is also a notable peak for the left bank at around 10 km (Figure 7c). This unexpected relation is most likely to be the result of bank protection centred on this location.

The coefficients of the fitted model are shown in Figures 5b (left bank) and 6b (right bank). The plots for composite banks with 30% gravel and composite banks with 60% gravel are similar. These two binary variables are competing at positions less than 30 km upstream and, therefore, convey similar information. Bedrock is important upstream of 30 km as might be expected. Cohesive banks have some influence between 20 km and 30 km upstream while banks with 100% gravel, which rarely occurs, has limited influence.

Meander

Figures 5a, 7 (left bank), 6a, and 8 (right bank) demonstrate that, as for history, whether the channel is meandering or not appears to have limited influence on bank erosion. Examining the deviance ratios closely reveals that this small influence is greatest between 10 km and 30 km upstream. The coefficients for meandering relate closely to the deviance ratios. In particular, two areas with large negative coefficients occur at around 10 km (and 16 km, right bank) and 24 km (and 28 km, right bank) upstream. As for material (above), the relation observed at around 10 km is likely to be the result of bank protection. The relation observed at 24–28 km may be explained by the controlling influence of bank material, in this case bedrock that outcrops where the meander cuts into the valley side. Overall, the Afon Dyfi has relatively few meandering sections, and in many cases these have been the focus for bank protection measures. In the headwater reaches, meanders are often confined by cohesive glacial drift or bedrock. Together these reduce the expected magnitude of correlation between meandering and the presence of bank erosion.

Stream Power

The most notable feature of the plots of the coefficients and deviance ratios of the GWLR for stream power against position upstream (Figure 5a, 6a, 7, and 8) is that the influence is greatest for the upper reaches (around 23–27 km to 34 km upstream). However, interestingly the relation is inverse for these positions, which was not anticipated. This can be explained in terms of the large values of unit stream power for the upper reaches. Where stream power is very high, there is little erodible material present. Where stream power is lower, the material present includes glacial terraces that are more likely to be eroded than the bedrock. In the lower reaches, an inverse relation is present for several positions. These relations arise in sections of the river where gravel bars result in both relatively low stream power (since channel width is maximised) and bank erosion. It is necessary to remember that the values of bankfull stream power are all above the critical threshold for the median bed material. Also, bank protection, which has not been included in the model data, results in reaches of high stream power (hence the need for bank protection) and low erosion.

There are several positions where a direct relation between stream power and erosion results, the most important of which appears at 18 km upstream. For these locations, weaker alluvial material supports such a direct relation.

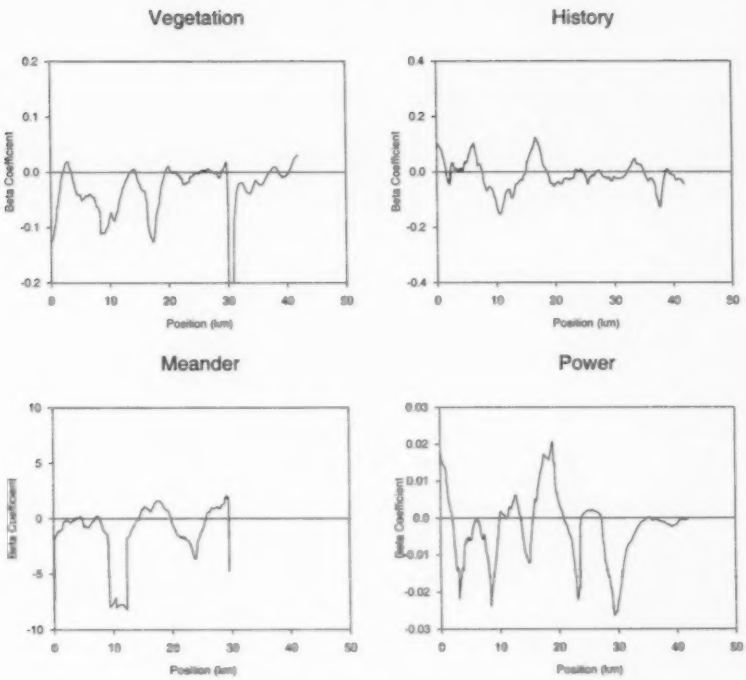

FIG. 5a. The Regression Coefficients for Vegetation, History, Meander, and Power Plotted Against Distance Upstream for the Left Bank.

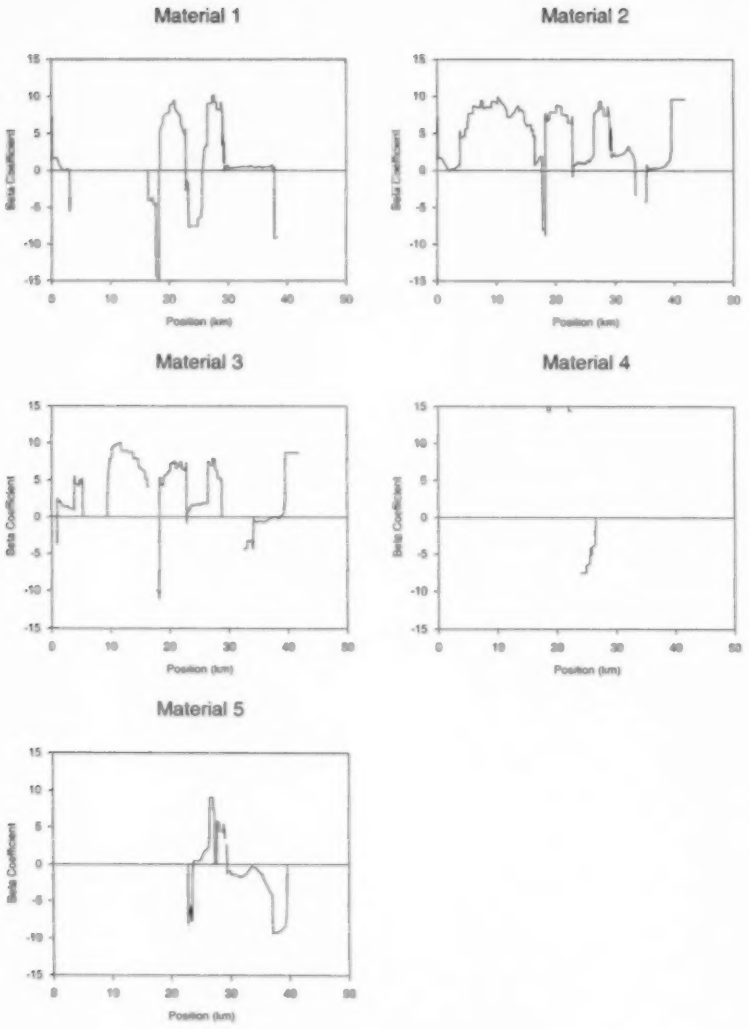

FIG. 5b. The Regression Coefficients for Material Plotted Against Distance Upstream for the Left Bank. Material is divided into five binary components as described in the text.

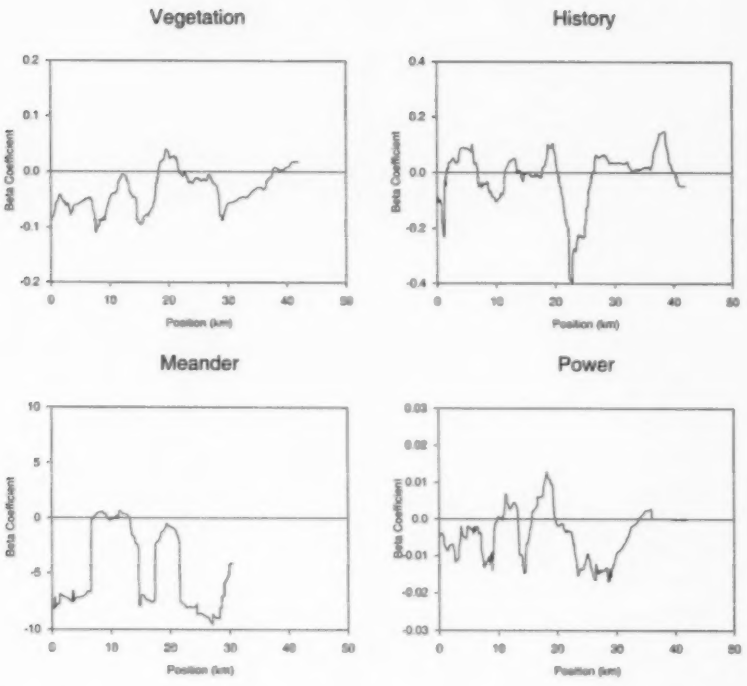

FIG. 6a. The Regression Coefficients for Vegetation, History, Meander, and Power Plotted Against Distance Upstream for the Right Bank.

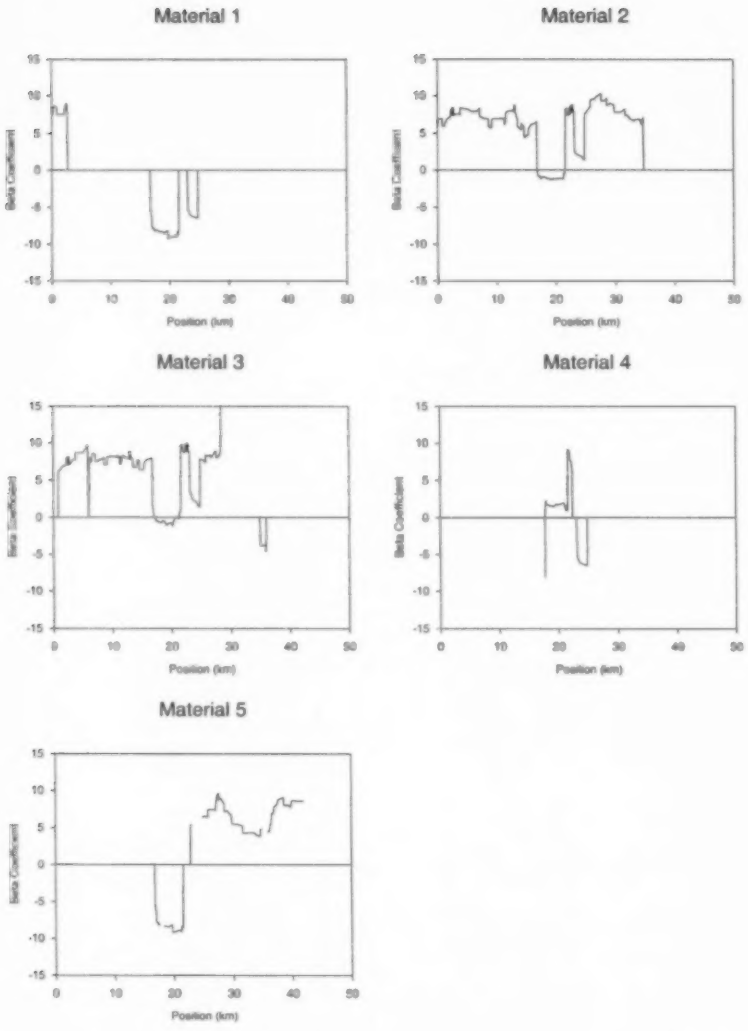

FIG. 6b. The Regression Coefficients for Material Plotted Against Distance Upstream for the Right Bank. Material is divided into five binary components as described in the text.

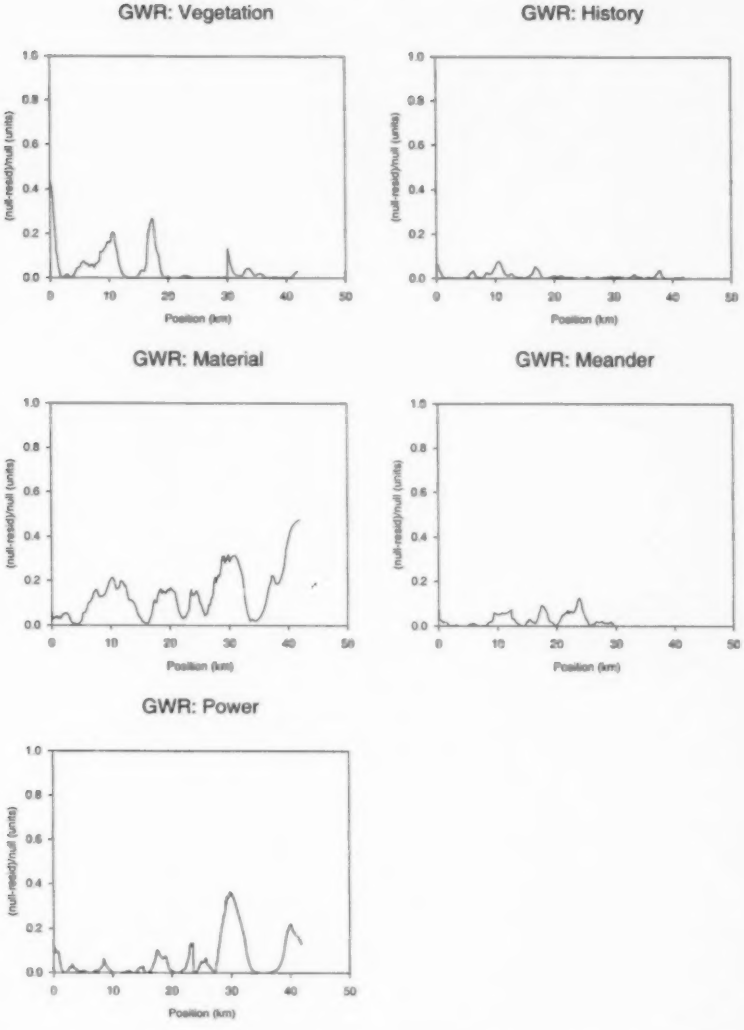

FIG. 7. Deviance Ratio ([Null deviance-Residual deviance]/Null deviance) Plotted Against Distance Upstream for the Left Bank.

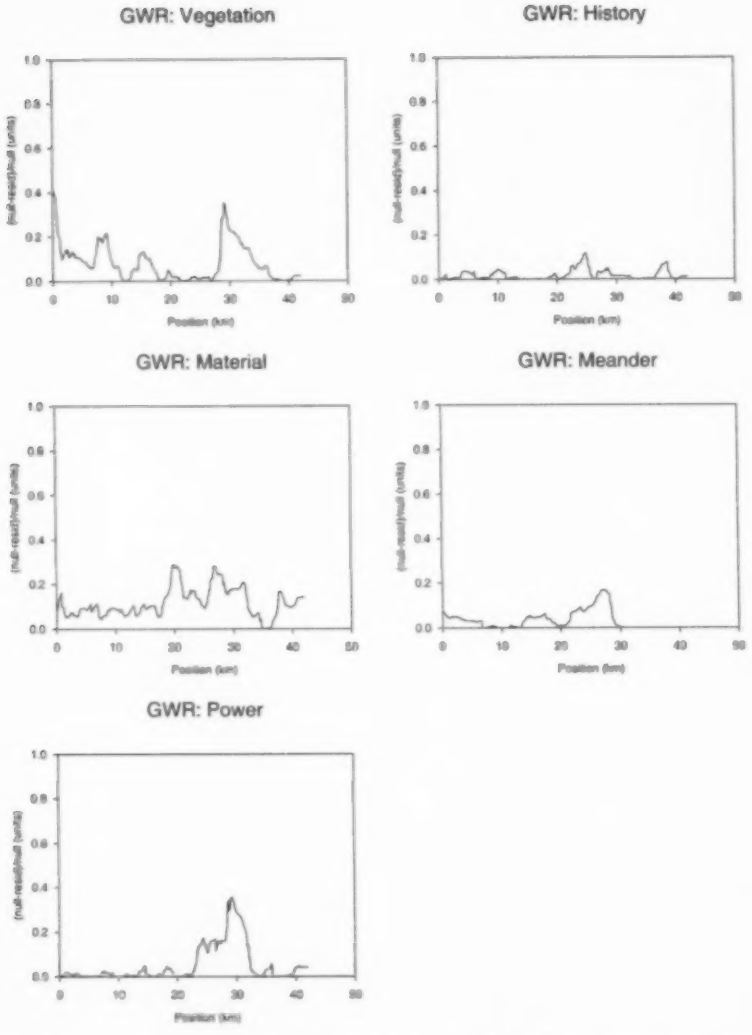

FIG. 8. Deviance Ratio ([Null deviance-Residual deviance]/Null deviance) Plotted Against Distance Upstream for the Right Bank.

DISCUSSION

GWLR in Physical Geography

GWR was developed primarily for application in human geography, for example, to socioeconomic data such as provided by the UK census. In human geography, it is reasonable to expect geographical variation in relations. For example, house prices in the urban area of Southampton might depend on an entirely different set of variables to house prices in its rural surroundings. One example would be the relation between house price and proximity to playing fields. In the urban area this might increase house price, whereas in the rural area it might decrease it. The most obvious measurable quantity that would describe such a difference is geographical location.

In physical geography, it is generally assumed that variation in some property over space cannot be due to space itself, but rather is due to some control that varies over space. The positivist view commonly held by physical geographers suggests that observations can be generalised into laws (Fotheringham, Brunsdon, and Charlton 2000). Thus, the goal of much research in physical geography is to find the underlying causal factors responsible for variation in some property (or in the present case, its relation with other properties) over space. In terms of regression modelling, it should be possible to build a single model that describes the entire set of possible circumstances that may arise. However, a general model describing such a large set of possible circumstances would include a large number of variables, possibly non-linear relations and a large number of interaction terms. Such a model would be cumbersome and difficult to interpret.

An advantage of GWR (or GWLR) is that by fitting the model locally, certain variables may be redundant in the local model because they are effectively held constant over space. Such terms have low-frequency or long-range variation. For example, in the case of the Afon Dyfi it is possible that elevation or some second-order derivative of elevation has the greatest influence on erosion. Since this property has not been measured, the *global* regression model would fail to describe well the relation between erosion and other geomorphological variables. However, within a region for which elevation varies little, the *local* model may explain most of the *local* variation in erosion. Thus, in an applied context, there is merit to suspending the search for causal factors, instead producing predictive models that have value locally and reduce the need to acquire large numbers of data in the field.

GWLR in Geomorphology

The attraction of GWLR analysis to geomorphology lies in the ability to explore spatial variation in the relations between driver variables and process—in this case, bank erosion. Such analysis permits the identification of zones of process control, which might be expected to vary between reaches of a river, but also between catchments. Using the Afon Dyfi as a case study, it becomes possible to comment on the widely held theoretical view that a specific sequence of process domains exist along a watercourse (Lawler 1992; Abernathy and Rutherford 1998). The GWLR analysis of the Dyfi data set demonstrates that the scale and sequencing of process domains may be much finer and more varied than existing models suggest. Rather, controls on bank erosion processes vary at a fine spatial frequency, primarily according to the available energy to drive erosion (stream power) and the resistance to that erosion provided by bank materials and vegetation. For the Dyfi at least, channel planform and the historical activity of the channel do not appear to influence the process or location of contemporary bank erosion directly, although both may control the location and nature of bank materials and vegetation.

GWLR for Management

The results from GWLR provide an insight into the local relations between erosion and the controls on that erosion. This information can be used, as part of a management strategy, to focus decisions regarding bank protection options. In a particular reach, the fact that, say, vegetation has an influence over erosion locally can be used to define the most suitable management recommendations *for that reach*. This allows managers to work with the natural fluvial system on a local basis. By providing understanding of the physical processes that create areas of erosion locally and of how these relations can change within a small area of the river, GWLR may allow river management to be more targeted and effective.

The remainder of this discussion is concerned with several issues which must be addressed when implementing GWLR.

Other Issues

Although some important advances of understanding and management application have been achieved, it is important to exercise caution when interpreting the results of GWLR, and this is particularly highlighted by the example of bank material. Since the categories of material exhibit relatively low frequency spatial variation, the successive peaks and troughs in the influence of material on erosion along the Dyfi may reflect local variation (large influence) and lack of local variation (small influence) between the classes of material. Thus, the deviance ratio will tend to be small where there is little or no variation in the category of material locally, and has the potential to be large only where variation in material exists locally.

An issue that is closely related to that above is significance. GLM is normally an exploratory technique: models are developed interactively, which provides an experienced user with a real "feel" for the data. This interactive procedure for fitting a logistic model sits awkwardly alongside GWR for which it is a requirement that fitting is automatic. In GWR, models must be calibrated for each desired location j of which there are likely to be between one hundred and one million, precluding interactive fitting. A consequence of automatic fitting is that some model coefficients are likely to be insignificant for certain locations, and for some locations it is possible that the entire model may not be significant. The issue of lack of significance is discussed by Brunsdon, Fotheringham, and Charlton (1996), but is effectively avoided when they map the entire set of m coefficients from GWR.

A further problem with GWLR that should be apparent from Equations 8 to 11 is how to choose an appropriate non-linear parameter a for the weighting function. In this paper, a choice was made following experimentation. However, an alternative is the cross-validation approach suggested by Cleveland (1979), in which the values at each i are omitted when calibrating the regression model at i:

$$\sum_{i=1}^{n} [y_i - y^*_{\neq i}(a)]^2 \tag{16}$$

where $y^*_{\neq i}(a)$ is the fitted value of y_i with the observation for location i omitted from the regression calibration. The optimal value of a can be chosen as the one that minimises the cross-validation score. An important point, however, is that where the objective is to explore relations (rather than to predict), variation in the regression model with a can be informative.

CONCLUSION

The major attraction of the GWLR approach lies in its ability to permit objective quantification of spatial variation in the magnitude and direction of the relations between some process-related response variable (bank erosion in this example) and its controlling variables. In the present example, for the Afon Dyfi, Wales, the model parameters and the residual deviance of the model varied greatly with distance upstream. The practical implication of this result is that different management practices should be implemented at different locations along the river. Thus, the approach presented allowed inference of spatially varying management practice as a consequence of spatially varying geomorphological process.

Future research will focus on extending the model to (i) the prediction of categories and magnitudes of erosion (c.f., a binary response), (ii) multivariate modelling, and (iii) new application areas in geomorphology.

LITERATURE CITED

Abernethy, B., and I. D. Rutherford (1998). Where along a River's Length Will Vegetation Most Effectively Stabilise Stream Banks? *Geomorphology* 23, 55–75.

Aitkin, M., D. Anderson, B. Francis, and J. Hinde (1989). *Statistical Modelling in GLIM*. Oxford: Oxford University Press.

Atkinson, P. M., and R. Massari (1998). Predicting the Relative Likelihood of Landsliding in the Central Appennines, Italy. *Computers and Geosciences* 24, 373–85.

Augustin, N. H., M. A. Mugglestone, and S. T. Buckland (1996). An Autologistic Model for the Spatial Distribution of Wildlife. *Applied Ecology* 33, 339–47.

Bagnold, R. A. (1966). An Approach to the Sediment Transport Problem from General Physics; *U.S. Geol. Survey Prof. Paper* 422-J.

Baker, R. J., and J. A. Nelder (1978). *The GLIM System: Release 3*. Oxford: Numerical Algorithms Group.

Brunsdon, C., A. S. Fotheringham, and M. E. Charlton (1996). Geographically Weighted Regression: a Method for Exploring Spatial Nonstationarity. *Geographical Analysis* 28, 281–98.

Bull, L. J. (1997). Magnitude and Variation in the Contribution of Bank Erosion to the Suspended Sediment Load of the River Severn. *Earth Surface Processes and Landforms* 22, 1109–1123.

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association* 74, 829–36.

Collett, D. (1991). *Modelling Binary Data*. London: Chapman and Hall.

Dobson, A. J. (1990). *An Introduction to Generalised Linear Models*. London: Chapman and Hall.

Fotheringham, S., C. Brunsdon, and M. Charlton (2000). *Quantitative Geography. Perspectives on Spatial Analysis*. London: Sage.

German, S. E. (2000). A Geomorphological Approach to the Strategic Management of River Bank Erosion: A Case Study of the Afon Dyfi. Ph.D. diss., University of Southampton.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. New York: Oxford University Press.

Gurnell, A. M. (1997). Channel Change on the River Dee Meanders, 1946–1992, from the Analysis of Air Photographs. *Regulated Rivers: Research and Management* 13, 13–26.

Healy, M. J. R. (1988). *GLIM: An Introduction*. Oxford: Clarendon Press.

Hooke, J. M. (1995). Processes of Planform Change on Meandering Channels in the UK. In *Changing River Channels*, edited by A. M. Gurnell and G. E. Petts, 87–115. Chichester: John Wiley and Sons Ltd.

Jones, A. P. (1995). Provenance and Palaeoenvironments of Late Quaternary Sediments and Their Associated Landforms. Ph.D. diss., University of Wales, Aberystwyth.

Lawler, D. M. (1992). Process Dominance in Bank Erosion Systems. In *Lowland Floodplain Rivers: Geomorphological Perspective*, edited by P. A. Carling and G. E. Petts. Chichester: John Wiley and Sons Ltd.

Lawler, D. M. (1993). The Measurement of River Bank Erosion and Lateral Change: A Review. *Earth Surface Processes and Landforms* 18, 777–821.

Lawler, D. M., C. R. Thorne, and J. M. Hooke (1997). Bank Erosion and Instability. In *Applied Fluvial Geomorphology for River Engineering and Management* edited by C. R. Thorne, R. D. Hey, and M. D. Newson, 137–72. Chichester: John Wiley and Sons Ltd.

Matheron, G. (1965). *Les Variables RégionalisÉes et Leur Estimation*. Paris: Masson.

Matheron, G. (1971). *The Theory of Regionalized Variables and its Applications*. Fontainebleau: Centre de Morphologie MathÉmatique de Fontainebleau.

Millar, R. G. (2000). Influence of Bank Vegetation on Alluvial Channel Patterns. *Water Resources Research* 36, 1109–18.

Rhoads, B. L. (1991). A Continuously-Varying Parameter Model of Downstream Hydraulic Geometry. *Water Resources Research* 27, 1865–72.

Rhoads, B. L. (1992). Statistical Models of Fluvial Systems. *Geomorphology* 5, 433–55.

Sear, D. A., M. D. Newson, and A. Brookes (1995). Sediment Related River Maintenance: The Role of Fluvial Geomorphology. *Earth Surface Processes and Landforms* 20, 629–47.

Simon, A., and C. R. Hupp (1992). Geomorphic and Vegetative Recovery Processes along Modified Stream Channels of West Tennessee. *U.S. Geological Survey Open-File Report*. Nashville, Tenn.: USGS, 91–502.

Simon, A., and S. Darby (1995). Preliminary Analysis of Bank-Stability Problems on the Upper Missouri River below Ft. Peck Dam. *Report to the USDA-Agricultural Research Service*. Oxford, Miss.: USD National Sedimentation Laboratory.

Thorne, C. R. (1982). Processes and Mechanisms of River Bank Erosion, in *Gravel-bed Rivers*, edited by R. D. Hey, J. C. Bathurst, and C. R. Thorne, 227–71. Chichester: John Wiley and Sons.

Thorne, C. R. (1990). Effects of Vegetation on Riverbank Erosion and Stability. In *Vegetation and Erosion*, edited by J. B. Thornes, 125–43. Chichester: John Wiley and Sons Ltd.

Thorne, C. R. (1997). Channel Types and Morphological Classification. In *Applied Fluvial Geomorphology for River Engineering and Management* edited by C. R. Thorne, R. D. Hey, and M. D. Newson, 175–222. Chichester: John Wiley and Sons.

Thorne, C. R., R. G. Allen, and A. Simon (1996). Geomorphological River Channel Reconnaissance for River Analysis, Engineering and Management. *Transactions of the Institute of British Geographers* 21, 469–83.

Uno, T., T. Sugii, M. Hayashi (1994). Logit Model for River Levee Stability Evaluation Considering the Flood Return Period. *Structural Safety* 14, 81–102.

Wang S.-Q., and D. J. Unwin (1992). Modelling Landslide Distribution on Loess Soils in China: An Investigation. *International Journal of Geographical Information Systems* 6, 391–405.

Wharton, G. (1995). Information Form Channel Geometry-Discharge Relations. In *Changing River Channels*, edited by A. Gurnell and G. Petts, 325–46. Chichester: John Wiley and Sons Ltd.

Richard L. Church
James R. Marston

Measuring Accessibility for People with a Disability

This paper discusses some of the inherent problems associated with measuring accessibility for people on a landscape of surfaces, barriers, and travel modes. Along with this discussion we propose a new perspective for measuring accessibility with a focus on people with differing abilities. Even though our focus is on people with a physical disability, such an approach can be easily extended and is able to be generalized to other needs and differences. Traditional measurements of accessibility are flawed, as they fail to directly account for mobility and physical differences among people. They ignore structural barriers and individual mobility limitations that affect travel time, effort, and even successful completion. To make sense of this dilemma, we propose an accessibility measurement framework that includes measures of absolute access, gross access, closest assignment access, single and multiple activity access, probabilistic choice access, and relative access. Most of these measures of access have been proposed by others, but our framework attempts to codify an approach that helps to overcome weaknesses in using only the absolute access measurement currently used in ADA compliance. Such measures can be used to map accessibility as well as to help select the mitigation or renovation projects that yield the greatest increase in accessibility for people with disabilities. We argue that for many urban and building design problems providing absolute access for people with physical disabilities should be accompanied by the use of a relative access measurement, so that removing barriers can be done in the order that provides the greatest improvement in access for a given level of expenditure.

1. INTRODUCTION

Accessibility is an important characteristic of the geography of space, whether it involves a small area (e.g., elements within a building) or a large region (e.g., elements within a metropolitan area). It is frequently included as a goal in transportation planning, land use planning, and building design. The reason for this is that most would argue that accessibility is something to value and improve in an urban setting. Beyond acknowledging the importance of accessibility as a goal in planning, it has seldom

Richard Church is a professor of geography in the Department of Geography, University of California at Santa Barbara. E-mail: church@geog.ucsb.edu. *James Marston is a postdoctoral researcher in the Department of Geography, University of California at Santa Barbara. E-mail:* marstonj@geog.ucsb.edu.

Geographical Analysis, Vol. 35, No. 1 (January 2003) The Ohio State University
Submitted: February 21, 2002. Revised version accepted: July 21, 2002.

been translated into performance measures by which policies are evaluated (Handy and Niemeier 1997). Without an accepted measurement approach for accessibility in a planning problem, it is difficult if not impossible to compare alternatives in a rational manner with respect to changes in accessibility. Perhaps Gould (1969) stated it best when he said: "accessibility is.... a slippery notion.... One of those common terms that everyone uses until faced with the problem of defining and measuring it." There have been some notable exceptions to the paucity of using a measurable accessibility criterion in public policy. One of these is the definition of accessibility for people with disabilities in their approach to a building. The American with Disabilities Act (ADA) states in section 4.3.2 under the title of "Accessibility Guidelines for Buildings and Facilities" the following:

> At least one accessible route within the boundary of the site shall be provided from public transportation stops, accessible parking, and accessible passenger loading zones, and public streets or sidewalks to the accessible building entrance they serve. The accessible route shall, to the maximum extent feasible, coincide with the route for the general public

Measuring accessibility for ADA compliance, is then translated into a simple test by the policy: at least one accessible route to a building entrance must be provided from the street, from passenger loading zones and when appropriate from public transit stops and public parking. It is easy to test compliance for such a policy in that either appropriate access is provided or it is not. Thus, the public policy on access for people with disabilities boils down to: it is either provided within an absolute sense or it is not. This means that accessibility requirements within the context of the ADA rules are standards-based. It is important to recognize that the Code of Federal Regulations (28 CFR Part 36) dealing with the standards for accessible design describe how accessibility to specific elements (e.g., telephone, drinking fountain, or toilet facility) should be accomplished as well as the provision of accessible routes to such elements and building spaces. Not all elements need to be accessible, just a minimum required number. Thus, the measurement of accessibility is defined to the extent that it is easy to measure compliance. Although each element (e.g., toilet) is important, our objective here is to discuss the general concept of accessibility with respect to people with disabilities.

It should be acknowledged that the above policy and implied measurement approach (i.e., standards-based) has led to the improvement of building access in every city and town in the U.S. for people with physical impairments. Although this standards-based approach has been valuable, it lacks the sensitivity that other measures of accessibility might provide. Since the standard is to ensure that absolute access has been provided, little attention has been devoted to the value or quality of the access provided. Further, providing a second access route to a building is not given any value, by virtue that a first access route meets the standard. This paper focuses on the weakness of this standards-based approach without some attention being given to the value of relative access. Our objective is to present a paradigm for measuring access for those with physical or mobility impairments within an urban landscape, that extends beyond the standards-based approach. This proposed approach is sensitive to both the number of routes and the values of each access route provided. It is hoped that this proposed paradigm will help pave the way for future accessibility research as well as form the basis for which the current standards-based approach is enhanced in urban planning and architectural design. In the next section we present an overview of the literature on accessibility measures. In the subsequent section we begin the process of describing how such measures can be extended to people with a physical disability and enhance the standards-based approach of absolute access that is dictated by the ADA guidelines. Then, we propose several new measures of relative ac-

cess that we believe meet the spirit of the current law and provide guidance for improving urban design for people with disabilities. We demonstrate these concepts with several simple access issues faced by individuals on the University of California at Santa Barbara campus. We conclude with recommendations for future work.

2. MEASURING ACCESS, A REVIEW

Accessibility is determined by the spatial distribution of potential destinations, the ease of reaching each destination, and the magnitude, quality, and character of the activities found there (Handy and Niemeier 1997). The greater the number of potential destinations within some defined time or distance range, the greater the accessibility. The closer such choice destinations are within this maximum range, the higher the level of accessibility. An accessibility measure estimates the level of access to some type of activity from a starting location or home location to one or multiple locations of that activity given a travel mode, distance, time, and cost constraints. Weibull (1976) has presented an axiomatic framework for the development of accessibility measures. Conceptually, there are seven main types of measures for accessibility that have been the subject of a number of papers in the literature: 1) counting; 2) total sums of distances; 3) closest available; 4) gross interaction potential; 5) probabilistic choice, 6) net and maximum benefit; and (7) absolute. Albeit, there are other proposed accessibility measures, the above classes capture many of the elements of measuring access.

The simplest approach involves a count of the number of locations at which an activity can be found within some maximum distance, time or cost of travel from a given location or point i (Wachs and Kumagai 1973). Talen and Anselin (1998) call this a "container" measure. This means that the accessibility of activity k for a person at location i using travel type l can be calculated as:

$$A_{ikl} = \sum_{j \in M_{ikl}} O_{jk} \tag{1}$$

where:

A_{ikl} = the accessibility of person i or zone i with regard to activity k and travel type l

O_{jk} = the number of opportunities for activity k at location j.

d_{ijl} = the distance, travel time, or other measure of effort separating i and j for person of travel type l.

s_{kl} = the maximum distance or range over which an activity is considered accessible for travel type l. One can also use travel time or cost as a criterion as well

$M_{ikl} = \{j \mid d_{ijl} < s_{kl}\}$, the set of activity locations considered accessible.

Equation (1) represents a simple count of the total number of places where activity k is considered accessible in relation to location i and travel type l, within some maximum travel distance, time, or cost. The larger this value the greater number of opportunities are within reach of location i for activity k and the better the accessibility. We have chosen to differentiate possible differences in travel mode as well as individual by the index l, e.g., an older women who owns a car or a young adult who uses a wheelchair (see Hagerstrand (1970) for an introduction to time, space, and mobility constraints in travel; Miller (1991) and Kwan (1998) for a framework on measuring individual accessibility within a space-time framework; Hanson (1995), Hanson and

Schwab (1987), and Kwan (1999) for a discussion on travel and accessibility differences among people; and Marston, Golledge, and Costanzo (1997) who investigated the travel behavior for those who are vision impaired and rely on public transit). Pirie (1979) stated that zonal accessibility measures not only neglect the distribution of activities within a zone, but also assume that all individuals within a zone face the same set of opportunities. We use the index i to represent a location about which accessibility is to be measured, not a zone. If the index i is used to represent a zone, then the issues raised by Pirie are relevant.

The first measure of accessibility captures distance only in terms of some prespecified maximum distance, time, or cost of travel. It does not distinguish among the accessible sites of the activity to location i, in terms of distance. Ingram (1971) proposed a measure where accessibility at a given location i is defined as the sum of the spatial separations from all other locations j to that location. We can represent this measure as:

$$A_{ikl} = \sum_j d_{ijl} O_{jk} \tag{2}$$

In general, the larger the value, the less accessible are the activities for a given i. A central location among a distribution of points is more likely to have a lower accessibility value, and hence be more accessible than other points. If one imposed a maximum distance range for accessible sites (like in equation 1), then it would be difficult to compare values across the landscape, because a sum may be low (and hence, suggest great accessibility), when the sum only included 1 accessible site within the range as compared to the presence of many more available sites. Rather than sum over all possible activity locations, regardless of distance it makes sense to sum up through the m closest activity locations for activity k. This modified measure would represent a total sum of distances to the closest set of activity points. Then, a comparison of accessibility at various locations across a landscape would be meaningful.

In many public services, especially emergency services, accessibility is measured in terms of the closest available server or location. Access, then, is not based upon a gross number of opportunities and their proximity, or on personal choices, but on how far the closest server is from a given i. This accessibility measure has been used as a proxy or surrogate measure of effectiveness in the location of public services (Hodgart 1978). We can formalize this using the previous notation as:

$$A_{ikl} = d_{ikl} \tag{3}$$

where d_{ikl} is the closest location to i having a large enough amenity of activity k to serve a person at i with access type l. It should be recognized that the closest activity k to a location i may differ among those of different access types. This accessibility measure has been formalized in the p-median location problem where an activity is located on a network among p-sites out of n sites in order to maximize accessibility. (Hakimi 1964; ReVelle and Swain 1970). Whereas most other accessibility measures increase in value with increasing accessibility, this measure decreases in value as accessibility increases. Thus, when using this type of accessibility measure in urban design, the sense of optimization would be to minimize.

We know that people do not necessarily choose to go to their closest available activity when given a choice. The size or attractiveness of the activity site, the distance, and available travel modes all play a role in that choice. One of the models that has been used to distribute trips among various destinations is the gravity model. The gravity model has been used in a number of different types of applications to estimate

the interaction between one location and another. The gravity model estimates interaction in terms of attractiveness of an activity at a given location j (e.g., number of jobs, size of retail space, etc.) and distance. Hansen (1959) was the first to propose a gravity-based accessibility measure for some activity k, in the following manner:

$$A_{ikl} = \sum_{j \in M_{kl}} O_{jk} d_{ij}^{-\beta_l} \tag{4}$$

where:

β_l = empirical constant representing the inhibiting effects of distance (time or effort) on trip making for travel type l.

The above equation represents a gross measurement of accessibility for a person of type l starting at location or zone i with respect to some activity k. The higher the value of gross accessibility, the greater the access and number of opportunities for a given activity. Gross accessibility to a given activity (e.g., shopping for shoes) is the sum of all such activities within some maximum time, distance, or cost limit, each discounted by how distant each activity is away from location i. An activity location is counted high when it is close, and low when it is far away. This discount is based upon how distance or time affects the possibility of making a trip to that location for that activity. Pooler (1987, 1995) presents a critical discussion on this type of "potential" accessibility model. We have taken the liberty of extending the accessibility measure suggested by Hansen by adding subscripts for the type of activity (k) and the type of person (l). It makes sense to represent different types of people, based upon their mobility status (Hanson 1995; Marston, Golledge, and Costanzo 1997). For example, if a sub-population at a zone can only move about by bus, then their accessibility to a given activity k from zone i will be different than those who own cars and can use either the bus or their car. Examples of such differences are given in Hanson and Schwab (1987). From a more rigorous perspective, one can generate a model of accessibility related to (4) as:

$$A_{ikl} = \sum_{j \in M_{kl}} O_{jk} e^{-\alpha d_{ij}} \tag{5}$$

where the distance attenuating function is of an exponential form. Although the distance decay function may be expressed in many ways, the negative exponential form is probably the more widely accepted form used. In the remainder of this paper we will use the simple distance decay form of equation (4), even though the form in equation (5) is applicable as well.

People make choices in travelling and destination. The gross accessibility measure sums up all possible opportunities (discounted by a function of distance) over a range of distance, time, or cost. However, it is unlikely that each person actually visits each activity site. More than likely, a choice will be made. Perhaps people with mobility type l may choose to go to one of two malls for shoe shopping, even though there are several nearby stores. The intervening opportunities model attempts to capture different competing elements (like size and distance) in consumer choice (Stouffer 1940). Even trip chaining may make a more distant activity site the most suitable (Hanson 1980). The simplest form of measuring accessibility based upon probabilities can be done in the following way:

$$A_{ikl} = \sum_{j \in M_{ikl}} p_{ijkl} d_{ijl} \tag{6}$$

where p_{ijkl} = the probability that a person in zone i will go to zone j for activity k involving travel type l. This measure represents the average distance of travel to activity k as the distance to each site j times the probability that that site is the choice (Geertman and van Eck 1995). One approach to estimating the probability of selecting a given location j can be to divide the accessibility potential of that site by the sum of all of the accessibility potentials generated at i for an activity k.

$$p_{ijkl} = \frac{O_{jk} d_{ijl}^{-\beta_l}}{\sum_{r \in M_{ikl}} O_{rk} d_{irl}^{-\beta_l}} \tag{7}$$

Wang (2000) used equations like 6 and 7 to estimate job accessibility by transportation analysis zones in Chicago. We call this a probabilistic or choice-based accessibility measure. Since the sum of the probabilities over all j should equal 1, this model does not "double count" activities, but only accounts for those activity locations actually used. This at first may seem trivial, but consider the following real example. In an Agency for International Development project involving the location of health clinics in Colombia, Church (1979) discusses a situation where a clinic located in one village would not necessarily serve the inhabitants of a nearby village if there was a major cultural or political divide. A gross accessibility measure applied to the village would rightly consider it accessible. A choice accessibility model applied to the village would indeed capture the fact the nearby clinic would be "tabu."

It makes sense that individuals might travel based upon making those choices that maximize their net benefits, or simply put, maximize their consumer surplus. An accessibility measure for a given location i directed at some activity k could be a model that estimates the benefit associated with making the choice of location j that maximizes their benefit. Alternate structures are possible, like the sum of net benefits over all feasible choices (see, for example, Miller (1999) who provides a framework for measuring accessibility benefits within transportation networks).

Finally, it is important to note that accessibility can be measured in an absolute format. For example, access is either provided or it isn't. Either a building can be approached and entered by a person using a wheelchair or it can't. As stated earlier, absolute access is commonly used as a standards-based approach for measuring physical access. Either bathrooms have been retrofitted or built for wheelchair access or they haven't. As long as there is at least one accessible activity location for each activity type, for a given location i, absolute access has been provided and deemed acceptable. It makes sense, however, to use the other access measurements for people with disabilities (like closest available, gross access, and probabilistic choice access) in addition to absolute access as figures of merit for measuring the effectiveness of building retrofit plans, urban design, and new building layouts. For example, a ramp at the rear of a structure may provide absolute access, but it may make all people who need such a ramp to go out of their way as compared to a typical user. That is, absolute access doesn't capture the qualities of access measurements given above.

In the next section, we focus on the measurement of access for multiple activities. Although, all of the measures discussed in this section can be described within a multiple activity set, we will focus on gross-potential, probabilistic choice, and closest distance.

3. ACCESSIBILITY MEASURES FOR MULTIPLE ACTIVITIES

When measuring accessibility, it is important to look over the range of possible activities. That is, individuals don't just shop for shoes, but they go to the movies, visit the doctor, attend religious services, etc. They may have daily or weekly routines that represent a range of activities. The accessibility of various activities will differ, but for an individual, what is most meaningful is the totality of activity accessibility. To capture this consider the following equation of Total Gross Accessibility:

$$A_{il} = \sum_k \sum_j O_{jk} d_{ijl}^{-\beta_l} \tag{8}$$

where:

A_{il} = the total of Gross accessibility over all types of activities for an individual at i of type l.

Adding all accessibility measures for a given location and access type represents the grand total of all gross accessibility measures. There is the potential to have scaling problems as some activities are more important than others. To accommodate for this, we can weight each activity by some measure of importance. For example, let's consider the frequency of need (e.g., trips per year) for an activity as an importance weight for that activity. Thus, we can define Weighted Gross Accessibility as:

$$A_{il} = \sum_k f_{ikl} \sum_j O_{jk} d_{ijl}^{-\beta_l} \tag{9}$$

f_{ikl} = frequency of trips (e.g.,trips per year) for a person of type l located at i going for activity k.

This measure is appealing in that it estimates total accessibility for a location and type of access. Just as there can be different types of people in terms of mobility, there can be different types of people based upon the frequency of use of specific activities. For example, an older person may visit the doctor more often, go to school less often, and volunteer their time at agencies more often than a younger person. A person using a wheelchair may make less frequent or more frequent trips than another person with a different mode of access for a given activity. This is still simplistic, given that people make multi-stop trips and are confined by a space-time prism. Extensions to this simple gross estimate can be made by accounting for such behavior and constraints (e.g., Arentze, Borgers, and Timmermans 1994; Southworth 1985; Recker, Chen, and McNally 2001).

We can also extend the concept of total accessibility for a probabilistic choice model as well:

$$A_{il} = \sum_k f_{ikl} \sum_j p_{ijkl} d_{ijl} \tag{10}$$

This form also sums up accessibility over different types of activities, each weighted by the frequency of activity engagement. Again, trip-chaining behavior is not considered in this simple measurement.

Finally, within a public service context we can create a measure of total access for all activities assuming that each individual is served by their closest server of a given activity type, and weighting each activity by frequency of need or use. By taking the closest available activity access measure and summing over activities weighted by frequency, we get:

$$A_{il} = \sum_k f_{ikl} d_{ikl} \tag{11}$$

This measure represents total accessibility as weighted closest available service for each activity. This is often called weighted distance and is the form of accessibility that is used in the multi-type facility p-median problem (see the nested hierarchical and non-nested hierarchical p-median problems of Weaver and Church 1991).

4. RELATIVE ACCESSIBILITY

There are true differences in access based upon the type of individual, l, being addressed. Golledge (1993) discusses the idea that even when considering the exact same geographic space, people with differing abilities must use, access, and travel through that environment using different routes, such that the conception and use of that space is "transformed" for different users. He called for measures that would help explain how people with disabilities access obstacle-ridden space (Golledge, 1994). Let us consider the measurement of accessibility from an office in a university building to the closest coffee cart or café. Let's consider two people, one who uses a wheelchair and the other quite ambulatory, each leaving the same office (2843 Ellison Hall on the UCSB campus) for coffee. This second floor office is close to a stairwell that is on the most direct route to the closest coffee cart (see figure 1). The ambulatory person leaves the office, goes down one flight of stairs, and exits the building on the east entrance, bounds down a few more steps, and then makes a beeline for the coffee cart across the plaza, east of the building. The person using the wheelchair heads to the bank of elevators in the center part of the building, waits for an elevator, and then takes it to the first floor. That person exits the building through the front (south facing) doors, and then heads east around the building to the plaza, and then towards the coffee cart. Thus, they both have chosen and reached the same cart. The routes for both individuals are depicted in figure 1.

For both people the cart is accessible in terms of an absolute measure. In fact, each measure described above for a single activity accessibility measure could be used. For this example, it makes the most sense to apply the closest available access measure, i.e., the distance, time, or effort it takes to travel to the coffee cart from the office. For the ambulatory person, the route takes approximately 40 seconds. For the person using a wheelchair, the route is longer as it involves a less direct route and an elevator wait and ride. Such a trip is approximately 3 minutes and 30 seconds. The effort taken in terms of time spent is 5.25 times that of the ambulatory person. We can express the differences in accessibility in terms of the person using a wheelchair, l, relative to the ambulatory person, m, leaving adjacent offices (call it location i) in the following way:

$$R_{iklm} = \frac{d_{ikl}}{d_{ikm}} \tag{12}$$

where k represents the activity of getting a cup of coffee for this example, and

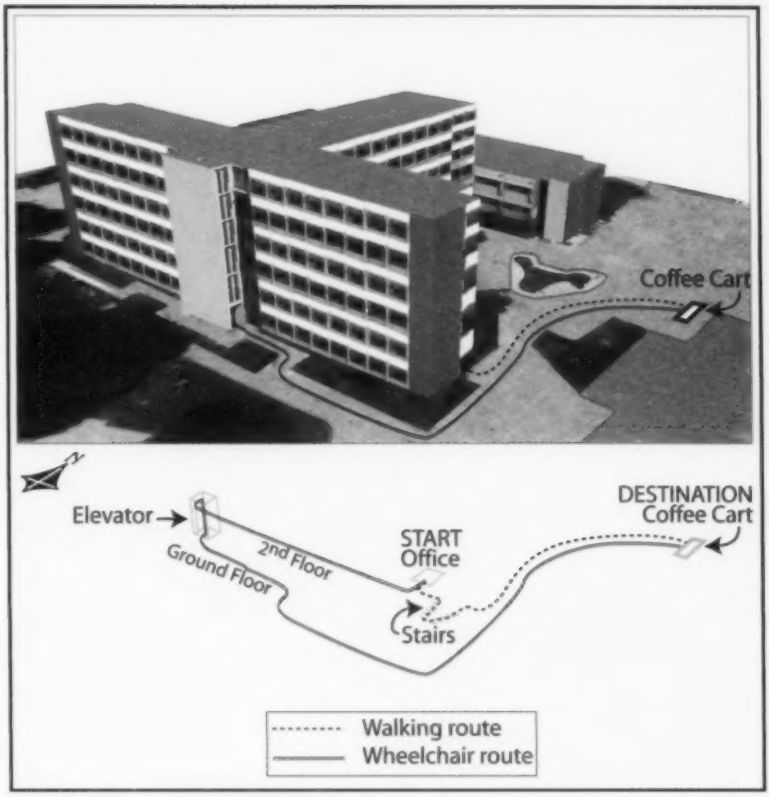

FIG. 1. Traveling between office 2843 in Ellison Hall and the nearest coffee cart on the UCSB campus, where one individual is ambulatory and the other uses a wheelchair.

R_{iklm} = relative accessibility of activity k from location i for person type l relative to person of type m.

If time is used then $R_{iklm} = 5.25$ for our example of getting coffee. Relative access is an important measure because it relates the differences in access relative to individual groups of users as well as it aids in understanding just what a physical impairment or constraint might represent in the effort needed to overcome obstacles in the environment (e.g., stairs and curbs) and travel. Relative access also helps to operationalize the space-time prism paradigm of Hagerstrand (1970) into an easy to understand and computable format. It is important to note that Lee and Lee (1998) have used the term relative access in a different way to represent the number of feasible competing routes to a destination and the number of mode changes. A similar notion, called route factor, was raised by Hay (1973) in analyzing the directness of routes between cities. De Jong and van Eck (1996) have used relative gross access and relative count access as access potentials and presented example maps for a region around Utrecht.

Marston and Golledge (2000) have used this type of time "penalty" to examine constraints faced by visually impaired travelers using their regular navigation methods. They also tested the effectiveness of using an auditory signage system, and compared travel time to those of a sighted user (Marston 2002). Subjects were tested in their relative ability in navigating and making five complex transfers between different transit modes in an urban transit environment.

Relative access can also be determined for a group or sum of activities. For example, let's assume that both people (in the above example) in performing their daily tasks during the work day, must travel three times to the photocopy machine (on the average), take two breaks that allow them to get a cup of coffee, travel several times to the bathroom, and attend several meetings within the building as well as go to lunch. Summing up accessibility over these types of activities and frequencies, we can obtain the following measure of relative access:

$$R_{iklm} = \frac{\sum_k f_{ikl} d_{ikl}}{\sum_k f_{ikl} d_{ikm}} \tag{13}$$

Where R_{iklm} = total relative access of person of type l compared to a person of type m, associated with location i assuming that person of type m has the same frequency of person l. Here relative access compares one class of individuals (access or mode type) to another based upon the same frequency of activities, but compounds the effects of spatial accessibility differences based upon the repetition of such activities

Relative access measures can be used to help understand the impacts of different design alternatives in buildings and public areas. For example, constructing a wheelchair ramp might provide absolute access to a building, but alternative placements of that ramp may differ considerably in terms of the impact on relative access. Whereas an absolute access measure can't distinguish between such positions, the relative access measure helps to measure the impact that the position of a ramp location has on the users. Suppose that a building already has an access ramp and is therefore accessible on an absolute scale and it also has at least one accessible restroom, water fountain, public telephone, etc. Then absolute access won't change if a second restroom is modified to be access friendly. But the gains in relative access might be considerable. That is, relative access is a good measure for determining the impact of barrier removal across the urban landscape or network.

Up to this point, we have introduced the concept of relative access within the context of closest assignment and weighted closest assignment. Actually, relative access can be defined for each of the access measures that have already been introduced. These include gross, probabilistic choice, closest assignment, and absolute as well as single, multiple, and weighted forms. When attempting to modify the landscape to improve access for some segment of the population (e.g., ADA compliance) it is important to provide absolute access. In fact, the implied measure that is used in access compliance is relative absolute access. The issue is to provide absolute access over various disabilities for all circumstances in which access is provided to those without disabilities. In essence, the objective of the ADA program has been to seek relative equivalence in terms of absolute access for various types of individual classes. Unfortunately, absolute access is a relatively crude form of access measurement and neglects the impact of the location of such barrier removals (e.g., lowering a public phone, making a restroom accessible, etc.) in terms of the movements required by those that use them. We argue that not only should access be a goal in terms of ADA compliance, but also that it be measured in terms of both absolute and relative access measures.

5. AN EXAMPLE IN MEASURING RELATIVE ACCESSIBILITY

The above concepts should be relatively understandable without an example. But, a real example encountered in campus planning at the University of California at Santa Barbara (UCSB) should help to underscore the importance of measuring relative accessibility. Even though we discuss this issue with respect to wheelchair users, we should point out that others pushing kids in strollers, paramedics pushing a gurney, courier personnel with a package cart, and campus audio-visual personnel pushing TV carts also encounter the same type of physical barriers in moving across campus. This access problem was presented to the campus ADA committee by one of the authors (Marston) in 1998. The analysis of relative penalty supported arguments made to make several improvements for the routes in question (i.e., an additional safety island for crossing the bikeway and making a route smooth so that it could accommodate those using wheelchairs). Figure 2 gives the layout of the central "core" of the UCSB campus. The University Center, which houses the bookstore and a number of food venders, is located in the bottom right portion of the figure. The Women's Center is located north of the University Center, close to building 477. There are several possible routes that are given in the figure that can be taken in going from the University Center to the Women's Center. A typical person would walk "Route A," which heads directly north along the east edge of Storke Plaza, crosses a bikeway with safety islands, and then makes a slight jog to the entrance to the Women's Center. The distance of this route is 650 feet. This heavily traveled route offers a bike crossing that splits the two lanes of traffic and offers a safety island so that people cross only one-way bike traffic at a time. Route A is the shortest walkable route and represents the route chosen by those who are relatively ambulatory.

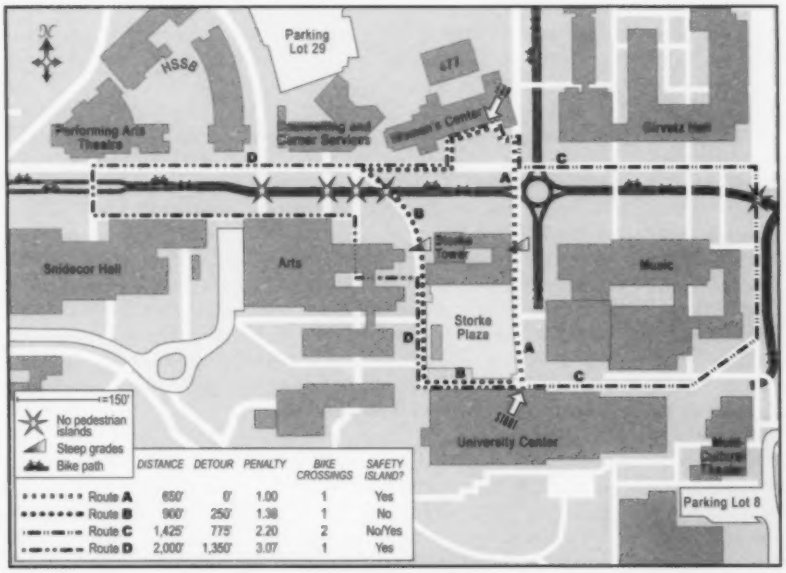

FIG. 2. Comparing Routes and Access Between the Women's Center and The University Center.

Unfortunately, the slope of the sidewalk along the east side of Storke Plaza is too steep for many people using wheelchairs. Even though we give this example associated with wheelchair users, people using a manual wheelchair generally must look for another route. One could travel along the west side of Storke Plaza, (Route B), then turn and travel towards the Counseling and Career Center. After crossing the bike way and reaching a major east-west oriented path, they could then head east to the Women's Center. This route is 900 feet or 250 feet longer than that of route A. In addition, the bike path at this point has no divided lanes with a safety island, making crossing more dangerous and difficult. The relative accessibility for having to take route B as compared to that of route A equals $\frac{900}{650} = 1.38$ (or 38% longer than route A). Even though this route is not too much longer than route A and does not have a steep slope, it had broken and cracked asphalt that made this route (Route B) impassable for those using wheelchairs. Until this area was repaved and smoothed, people using wheelchairs had to travel east from the University Center to the end of the Music building, turn north, continue heading north until crossing the bikepath, and then head west to the Women's Center. This route (Route C) adds 775 feet to the trip for a relative access measure (compared to those who are able to travel route A) of 2.19 (i.e., $\frac{1425}{650} = 2.19$) or 119% longer than route A. People using this route also have to cross the bikeway heading north, along the east side of the Music building, where there was no safety island, and then cross another bikeway close to the Women's Center. To avoid this very dangerous bikeway crossing along route C, a person could take Route D, which goes through the Arts building, heads west to Snidecor Hall, goes north across the divided bike path in front of the Performing Arts Theater, and then back east to the Women's Center. This route adds 1350 feet to the trip as compared to Route A, for a relative access penalty of 3.07, or 207% longer. Taking route D was considerably safer than route C. Until route B was repaved so that it was smooth and could allow wheelchair traffic, those people using wheelchairs faced a relative access penalty of 2.2 or 3.07 compared to those who are ambulatory. This level of penalty is not insignificant and can have other adverse effects on access to opportunities. For example, it has been reported to one of the authors that a student at a different university, who had a mobility impairment, took three years to gain her MA compared to the standard two for her cohort; some required classes could not be scheduled in the same term because class times were separated by only ten minutes of travel time and routes did not exist between the classes that could be traveled by her in that time constraint.

If a competitive route exists between an origin and destination that provides access to people with disabilities, then those people face no additional travel penalties as compared to a typical traveler. The greater the disparity between routes available to the ambulatory and those who have disabilities, the greater the difference in relative access. Measuring access only in terms of an absolute access measure does overlook the fact that significant disparities in access may remain between those with an impairment or limited mobility and the typical user.

An accessible route is supposed to coincide, when feasible, with the route for the general public. In many cases it is either not possible or it is cost prohibitive to make the route used by the general public entirely accessible to all people. We believe that the relative access measures proposed in this paper should be used to help make the most cost effective decisions regarding the selection of routes between buildings, like those on a university campus, that should be made accessible when removing barriers.

Hagerstrand (1970) introduced the notions of authority constraints and coupling constraints in describing the concept of the space-time prism. Coupling constraints

involve activities where more than one individual meets for an activity and authority constraints involve restrictions on access due to some social, political, or legal restriction on access. Measuring access for people of differing abilities must account for such conditions as well. For example, the scheduling constraints mentioned above involving the person with a mobility restriction hampered that individual from taking specific classes in the same term. Thus, authority conditions (e.g., published class schedules) can have a significant impact on relative accessibility. Forer and Huisman (2000) and Huisman and Forer (1998a, 1998b) discuss space-time accessibility issues within the construct of sequencing and schedules. Including sequencing and scheduling conditions will, in general, exacerbate problems of relative accessibility between people with differing abilities. Just as removing physical barriers to open up new routes of accessibility, it may also be possible to remove "schedule barriers" and improve absolute and relative accessibility by making changes to authority constraints (e.g., change published course schedules).

It is important to recognize that relative accessibility may differ greatly among differing spatial scales. For example, accessibility on a university campus differs from that of accessibility within a city. As overall travel distances increase for the ambulatory, differences in relative accessibility for those with differing abilities will likely decrease. Unfortunately, many daily activities (e.g., going to lunch or using a restroom) involve short trips for the ambulatory, but possibly time consuming trips for those with a disability.

6. CONCLUSIONS

Providing equal access to all is the goal of the ADA. This act has helped improve access to a variety of facilities, including public sector facilities such as parks, libraries, and courthouses and private sector facilities such as stores, motels, and theaters. Even the availability of accessible restrooms plays a role in making the environment of a facility useful. We know that disparities in access will continue to exist, but as money is spent on renovation, remodeling, and removing barriers across the urban landscape, attention should be directed towards making cost effective decisions, decisions that will help make the greatest improvement in overall accessibility. We have proposed that a relative access measure be used in conjunction with the more traditional measure of absolute access to help make such decisions. We have given several examples that show the relative penalty in comparing those using a wheelchair to an ambulatory person. We hope that this discussion and proposal will lead to further research as well as better decision-making.

LITERATURE CITED

Arentze, T. A., W. J. Borgers, and H. J. P. Timmermans (1994). Multistop-Based Measurements of Accessibility in a GIS Environment. *International Journal of Geographical Information Systems* 8, 343–356.

Church, R. L. (1979). An Introduction and Guide to the Site Location of Public Facilities Utilizing the 'Gas Program.' Appendix C to Final Report to AID, Contract AID/TA-C-1363, Lyndon B. Johnson School of Public Affairs, The University of Texas at Austin.

De Jong, T., and J. R. van Eck (1996). Location profile-based measures as an improvement on accessibility modeling in GIS. *Computers, Environment, and Urban Systems* 20, 181–190.

Forer, P., and O. Huisman (2000). Space, time and sequencing: substitution at the physical/virtual interface. In *Information, Place, and Cyberspace: Issues in Accessibility* edited by D. G. Janelle and D. C. Hodge. Berlin: Springer.

Geertman, S. C., and J. R. van Eck (1995). GIS and Models of Accessibility Potential, An Application in Planning. *International Journal of Geographical Information Systems* 9, 67–80.

Golledge, R. G. (1993). Geography and the Disabled—a Survey With Special Reference to Vision Impaired and Blind Populations. *Transactions of the Institute of British Geographers* 18, 63–85

Golledge, R. G. (1994). Disability, Barriers And Discrimination. International Conference on The New Distributional Ethics, Differentiation and Discrimination, Gothenburg, Sweden.

Hagerstrand, T. (1970). What About People in Regional Science? *Papers of the Regional Science Association* 24, 7–21.

Hakimi, S. (1964). Optimum Location of Switching Centers and the Absolute Medians of a Graph. *Operations Research* 12, 450–459.

Handy, S. L., and D. A. Niemeier (1997). Measuring Accessibility, An Exploration of Issues and Alternatives. *Environment and Planning* A 29, 1175–1194.

Hansen, W. G. (1959). How Accessibility Shapes Land Use. *Journal of the American Institute of Planners* 25, 73–76.

Hanson, S. (1980). Spatial Diversification and Multipurpose Travel: Implications for Choice Theory. *Geographical Analysis* 12, 245–257.

Hanson, S. (1995). Getting There, Urban Transportation in Context. In *The Geography of Urban Transportation* edited by S. Hanson, pp. 3–25. New York: Guilford.

Hanson, S., and M. Schwab (1987). Accessibility and Intraurban Travel. *Environment and Planning* A, 735–748.

Hay, A. (1973). *Transport For The Space Economy: A Geographical Study*. Seattle: University of Washington Press.

Hodgart R. L. (1978). Optimizing Access to Public Services. *Progress in Human Geography* 2, 17–48.

Huisman, O., and P. Forer (1998a). Towards a Geometric Framework for Modeling Space-Time Opportunities and Interaction Potential. Presented at the International Geophysical Union Commission on Modeling Geographical Systems meeting, Lisbon, Portugal.

Huisman, O., and P. Forer (1998b). Computational Agents and Urban Life Spaces: A Preliminary Realization of the Time-Geography of Student Lifestyles. Proceedings of the 3rd International Conference on GeoComputation, University of Bristol (http://www.geocomputation.org/1998/68/gc_68a.htm).

Ingram, D. R. (1971). The Concept of Accessibility, A Search for an Operational Form. *Regional Studies* 5, 101–117.

Kwan, M. P. (1998). Space-Time and Integrated Measures of Individual Accessibility, A Comprehensive Analysis Using a Point-Based Framework. *Geographical Analysis* 30, 191–216.

Kwan, M. P. (1999). Gender and Individual Access to Urban Opportunities, A Study Using Space-Time Measures. *Professional Geographer* 51, 210–227.

Lee, K. and H-Y Lee (1998). A New Algorithm for Graph-Theoretic Nodal Accessibility Measurement. *Geographical Analysis* 30, 1–14.

Marston, J. R. (2002). Towards an accessible city: empirical measurement and modeling of access to urban opportunities for those with vision impairments, using Remote Infrared Audible Signage. Unpublished dissertation, University of California, Santa Barbara.

Marston, J. R., R. G. Golledge, and M. Costanzo (1997). Investigating Travel Behavior pf Nondriving Blind and Vision Impaired People, The Role of Public Transit. *The Professional Geographer* 49, 235–245.

Marston, J. R., and R. G. Golledge (2000). Towards an Accessible City, Removing Functional Barriers to Independent travel for Blind and Vision Impaired, A Case for Auditory Signs. Final Report, University of California Berkeley, University of California Transportation Center, Grant # UCTC 65V430

Miller, H. J. (1999). Measuring Space-Time Accessibility Benefits Within Transportation Networks, Basic Theory and Computational Procedures. *Geographical Analysis* 31, 1–26.

Miller, H. J. (1991). Modeling Accessibility Using Space-Time Prism Concepts Within Geographic Information Systems. *International Journal of Geographical Information Systems* 5, 287–301.

Pirie, G. H. (1979). Measuring Accessibility, A Review and Proposal. *Environment and Planning* A 11, 299–312.

Pooler, J. (1987). Measuring Geographical Accessibility, A Review of Current Approaches and Problems in the Use of Population Potentials. *Geoforum* 18, 269–289.

Pooler, J. A. (1995). The Use of Spatial Separation in the Measurement of Transportation Accessibility. *Transportation Research* A 29, 421–427.

Recker, W. W., C. Chen, and M. G. McNally (2001). Measuring the Impact of Efficient Household Travel Decisions on Potential Travel Time Savings and Accessibility Gains. *Transportation Research Part A* 35, 339–369.

ReVelle, C. S., and R. W. Swain (1970). Central Facilities Location. *Geographical Analysis* 2, 30–42.

Southworth, F. (1985). Multi-Destination, Multi-Purpose Trip Chaining and Its Implications for Locational Accessibility, A Simulation Approach. *Papers of the Regional Science Association* 57, 107–123.

Stouffer, S. (1940). Intervening Opportunities, A Theory Relating Mobility and Distance. *American Sociological Review* 5, 845–867.

Talen, E., and L. Anselin (1998). Assessing Spatial Equity, An Evaluation of Measures of Accessibility to Public Playgrounds. *Environment and Planning* A 30, 595–613.

Wachs, M., and T. G. Kumagai (1973). Physical Accessibility as a Social Indicator. *Socio-Economic Planning Sciences* 7, 437–456.

Wang, F. (2000). Modeling Commuting Patterns in Chicago in a GIS Environment, A Job Accessibility Perspective. *Professional Geographer* 52, 120–133.

Weaver, G. R., and R. L. Church (1991). The Nested Hierarchical Median Facility Location Problem. *INFOR* 29, 100–115.

Wiebull, J. W. (1976). An Axiomatic Approach to the Measurement of Accessibility. *Regional Science and Urban Economics* 6, 357–379.

Instructions to Contributors

Manuscripts should be submitted to Alan Murray, Department of Geography, The Ohio State University, 1036 Derby Hall, Columbus, Ohio 43210-1361, USA (email: murray.308@osu.edu). Electronic submissions (Microsoft Word or PDF format as one complete file) or four (4) printed copies should be sent to the editor at the above address. Manuscripts and illustrations should conform to the stylistic requirements set forth in these instructions. Note that manuscripts longer than twenty-five (25) printed pages may not be considered for publication.

Books for review should be sent to the editor at the address above.

Format. Manuscripts may be sent via mail, double-spaced, on one side only of white bond paper, 8 1/2 x 11 inches in size, with wide margins, especially on the left-hand side, to allow for editorial notations or submitted electronically following the same guidelines for size. English language only. Upon acceptance, contributors will be asked to supply an electronic copy in WordPerfect or Microsoft Word.

Abstract. An abstract, double-spaced, must accompany the article. It should summarize the main points of the paper.

Literature Cited. All works cited should be listed alphabetically by author name in a double-spaced list at the end of the manuscript. Bibliographic information should be ordered as follows: author last name, initials, publication date in parentheses, title of the work, and (in the case of books) the place of publication and publisher, or (in the case of articles) the periodical title, volume number, and inclusive page numbers. References in the text to the literature should be made by the last name(s) of the author(s) and year of publication enclosed in parentheses, along with any other pertinent information; e.g., (Smith 1979), (Smith 1979, p. 241), (Smith 1979, chap. 2).

Mathematics. Displayed equations must be identified by consecutive Arabic numbers in parentheses on the right. Expressions should be aligned and compound subscripts and superscripts clearly marked if there is any potential for confusion. The use of a lowercase "oh" for a zero subscript or superscript should be avoided. To avoid repetition, replace a long expression occurring several times with a symbol. The slash (/) should be used to separate the numerator and denominator of all intent fractions and short displayed fractions; complex displayed fractions may be built up. Negative powers should be substituted for compound fractions in complex expressions. Multiple dots and bars over expressions should be avoided where primes can be used. Indicate boldface characters and italic characters by using the respective attribute of Word or WordPerfect.

Authors Affiliation. Authors must supply their professional title, department, name of the institution, and email address.

Illustrations. Illustrations may be high quality, and executed on white paper or vellum, in black ink, with clear, medium weight, black lines and figures or may be submitted in JPEG or TIFF format. Formatted lettering should appear in illustrations. Figures should be provided at sizes no larger than 4 1/2 x 7 inches (full page) and preferably no larger that 4 1/2 x 3 1/2 inches (half page). They should be numbered consecutively, and the number and author's name should be penciled lightly on the back of each or the electronic file should indicate the number and author's name. All illustrations must have captions, which should not appear on the artwork but should be provided, double-spaced, on a page at the end of the manuscript. If there is any potential for doubt, the word top should be written on the back of the illustration or indicated on the electronic file.

Tables. Tables should be numbered consecutively throughout the article and provided on separate pages at the end of the manuscript. Each table must include a descriptive title and headings to columns.

Copyrighted Material. Permission to reprint any previously published material (e.g., tables, illustrations, and equations) must be obtained by the author from the copyright holder.

Copyright and Copying

Copyright © 2003 by The Ohio State University. All rights reserved. Article copying exclusively for personal or internal use, or for the personal use of specific clients, is permitted provided that the copy user or producer pay a fee of $2.00 per copy of any article, comment, book review, or other complete unit to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, Massachusetts 01923, USA (www.copyright.com). This limited permission does not extend to other kinds of copying, such as copying for general distribution, for resale, for creating new collective works, or for advertising and promotion purposes. Specific written permission for such copying must be obtained from The Ohio State University Press. Fee code: 0016-7363/03 $2.00 + 0.

An International Journal of
Theoretical Geography
ISSN 0016-7363

ARTICLES

Evaluating an Entropy-Maximizing Solution to the Ecological Inference Problem: Split-Ticket Voting in New Zealand 1999
Ron Johnston and Charles Pattie

Measuring Impacts of Condition Variables in Rule-Based Models of Space-Time Choice Behavior: Method and Empirical Illustration
Theo Arentze and Harry Timmermans

Urban Land Use for Transport Systems and City Shapes
Francesca Medda, Peter Nijkamp, and Piet Rietveld

Exploring the Relationship Between River Bank Erosion and Geomorphological Controls Using Geographically Weighted Logistic Regression
Peter M. Atkinson, S. E. German, D. A. Sear, and M. J. Clark

Measuring Accessibility for People with a Disability
Richard L. Church and James Marston